

# **Toward an Effective Short-Range Ensemble Forecast System**

**September 18, 2003**

**Tony Eckel and Cliff Mass**

**University of Washington Atmospheric Sciences Department**

**This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.**

# UW's Ensemble of Ensembles

Name	# of Members	EF Type	Initial Conditions	Forecast Model(s)	Forecast Cycle	Domain
<i>Homegrown</i>	ACME	SMMA	8 Ind. Analyses, 1 Centroid, 8 Mirrors	“Standard” MM5	00Z	36km, 12km
	ACME <sup>core</sup>	SMMA	Independent Analyses	“Standard” MM5	00Z	36km, 12km
	ACME <sup>core+</sup>	PMMA	“ “	8 MM5 variations	00Z	36km, 12km
<i>Imported</i>	PME	MMMA	“ “	8 “native” large-scale	00Z, 12Z	36km

**SMMA: Single Model Multi-Analysis**

**PMMA: Perturbed-model Multi-Analysis**

**MMMA: Multi-model Multi-Analysis**

**ACME: Analysis-Centroid Mirroring Ensemble**

**PME: Poor Man's Ensemble**

**MM5: PSU/NCAR Mesoscale Modeling System Version 5**

## “Native” Models/Analyses of the PME

	Abbreviation/Model/Source	Type	Resolution (~ @ 45 °N)		Objective Analysis
			Computational	Distributed	
	<b>avn</b> , Global Forecast System (GFS), National Centers for Environmental Prediction	Spectral	T254 / L64 ~55 km	1.0° / L14 ~80 km	SSI 3D Var
	<b>cmcg</b> , Global Environmental Multi-scale (GEM), Canadian Meteorological Centre	Finite Diff	0.9°×0.9°/L28 ~70 km	1.25° / L11 ~100 km	3D Var
	<b>eta</b> , limited-area mesoscale model, National Centers for Environmental Prediction	Finite Diff.	32 km / L45	90 km / L37	SSI 3D Var
	<b>gasp</b> , Global Analysis and Prediction model, Australian Bureau of Meteorology	Spectral	T239 / L29 ~60 km	1.0° / L11 ~80 km	3D Var
	<b>jma</b> , Global Spectral Model (GSM), Japan Meteorological Agency	Spectral	T106 / L21 ~135 km	1.25° / L13 ~100 km	OI
	<b>ngps</b> , Navy Operational Global Atmos. Pred. System, Fleet Numerical Meteorological & Oceanographic Cntr.	Spectral	T239 / L30 ~60 km	1.0° / L14 ~80 km	OI
	<b>tcwb</b> , Global Forecast System, Taiwan Central Weather Bureau	Spectral	T79 / L18 ~180 km	1.0° / L11 ~80 km	OI
	<b>ukmo</b> , Unified Model, United Kingdom Meteorological Office	Finite Diff.	5/6°×5/9°/L30 ~60 km	<i>same</i> / L12	3D Var

# Design of ACME<sup>core+</sup>

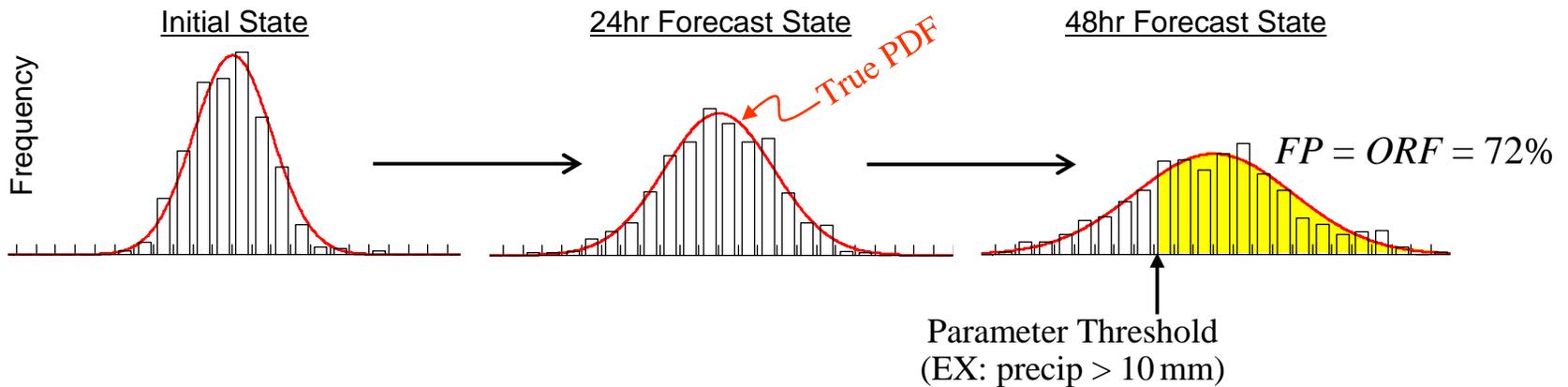
IC	ID#	PBL			Cloud Microphysics	Cumulus			Radiation	SST Perturbation	Land Use Table
			Soil	vertical diffusion		36km Domain	12km Domain	shallow cumulus			
ACME		MRF	5-Layer	Y	Simple Ice	Kain-Fritsch	Kain-Fritsch	N	cloud	<i>standard</i>	<i>standard</i>
<b>ACME<sup>core+</sup></b>											
avn	<b>plus01</b>	MRF	LSM	Y	Simple Ice	Kain-Fritsch	Kain-Fritsch	Y	RRTM	SST_pert01	LANDUSE.TBL.plus1
cmcg	<b>plus02</b>	MRF	5-Layer	Y	Reisner II (grpl), Skip4	Grell	Grell	N	cloud	SST_pert02	LANDUSE.TBL.plus2
eta	<b>plus03</b>	Eta	5-Layer	N	Goddard	Betts-Miller	Grell	Y	RRTM	SST_pert03	LANDUSE.TBL.plus3
gasp	<b>plus04</b>	MRF	LSM	Y	Shultz	Betts-Miller	Kain-Fritsch	N	RRTM	SST_pert04	LANDUSE.TBL.plus4
jma	<b>plus05</b>	Eta	LSM	N	Reisner II (grpl), Skip4	Kain-Fritsch	Kain-Fritsch	Y	cloud	SST_pert05	LANDUSE.TBL.plus5
ngps	<b>plus06</b>	Blackadar	5-Layer	Y	Shultz	Grell	Grell	N	RRTM	SST_pert06	LANDUSE.TBL.plus6
tcwb	<b>plus07</b>	Blackadar	5-Layer	Y	Goddard	Betts-Miller	Grell	Y	cloud	SST_pert07	LANDUSE.TBL.plus7
ukmo	<b>plus08</b>	Eta	LSM	N	Reisner I (mx-phs)	Kain-Fritsch	Kain-Fritsch	N	cloud	SST_pert08	LANDUSE.TBL.plus8
											<b>Perturbations to:</b> 1) <i>Moisture Availability</i> 2) <i>Albedo</i> 3) <i>Roughness Length</i>

Total possible combinations:

$$8 \times 5 \times 2 \times 2 \times 5 \times 3 \times 2 \times 2 \times 2 \times 8 \times 8 = \mathbf{1,228,800}$$

## Effects on EF Forecast Probability

- In a large, well-tuned EF, Forecast Probability ( $FP$ ) = Observed Relative Frequency ( $ORF$ )



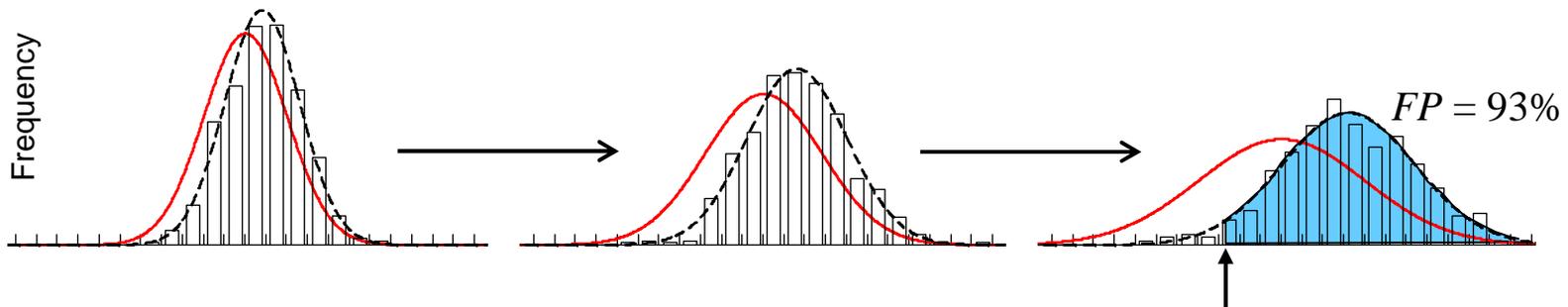
- In practice, things go awry from...

- Undersampling of the PDF (too few ensemble members)
- Poor representation of IC uncertainty
- **Model deficiencies**

1) **Model bias causes a shift in the estimated mean**

2) **Sharing of model errors between EF members leads to reduced variance**

- EF's estimated PDF does not match truth's PDF—degrades **resolution** and **reliability**



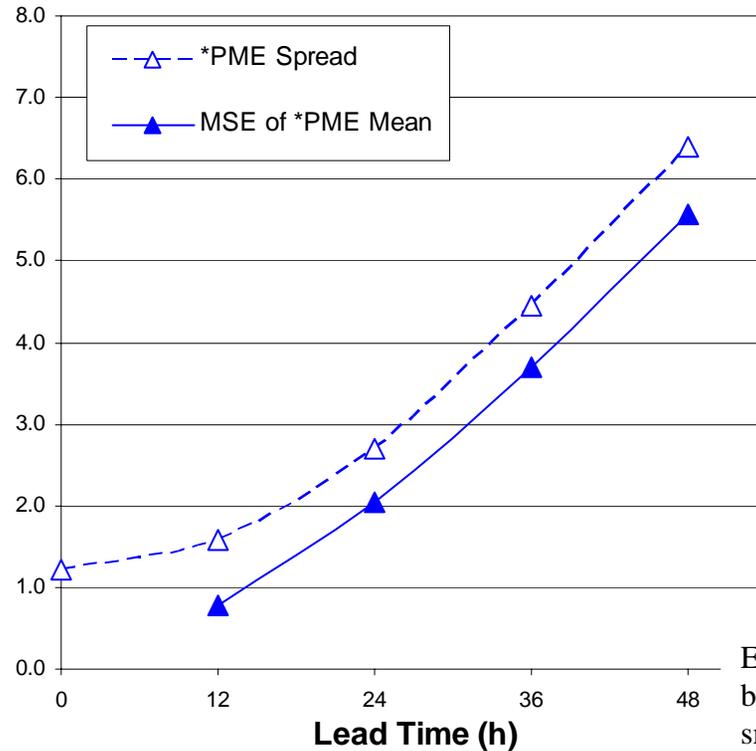
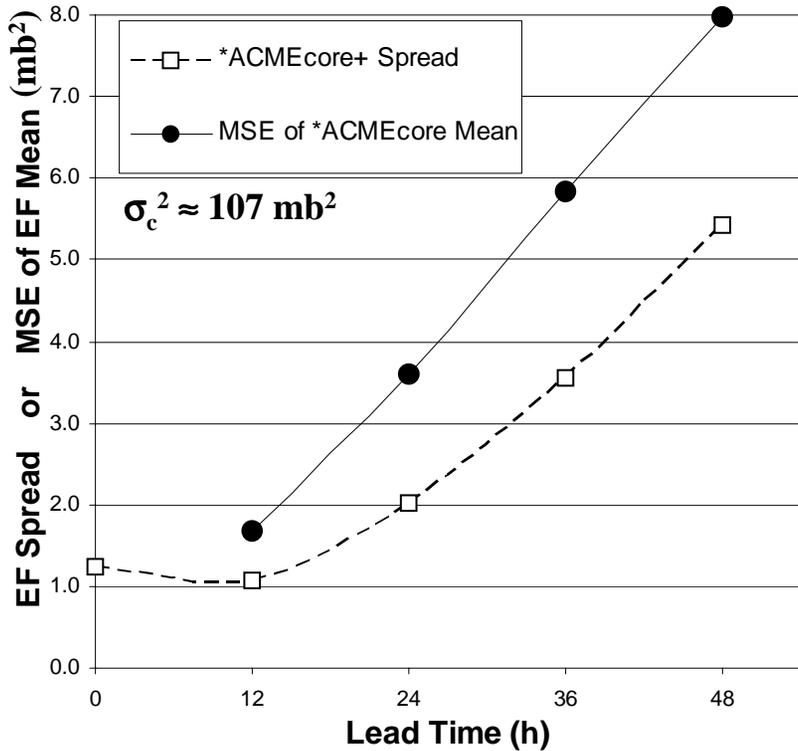
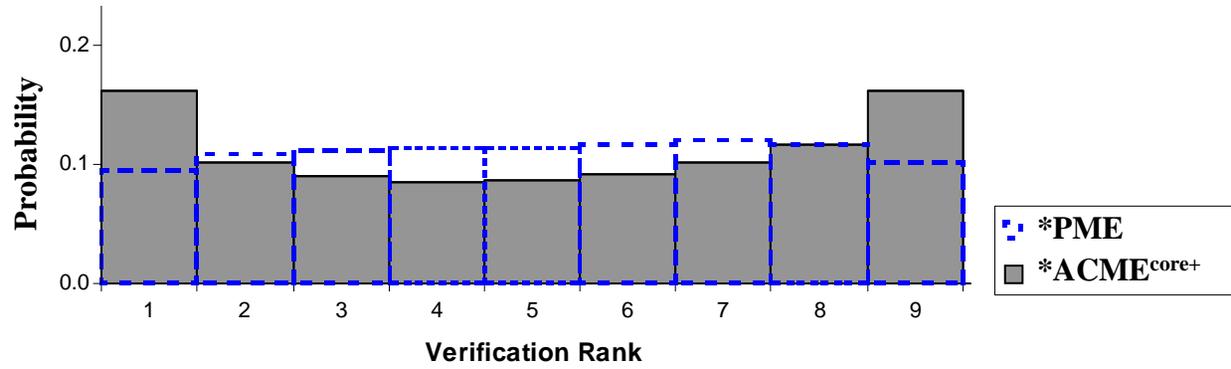
**Multimodel  
Vs.  
Perturbed-Model**

**PME  
Vs.  
ACME<sup>core+</sup>**



# Ensemble Dispersion

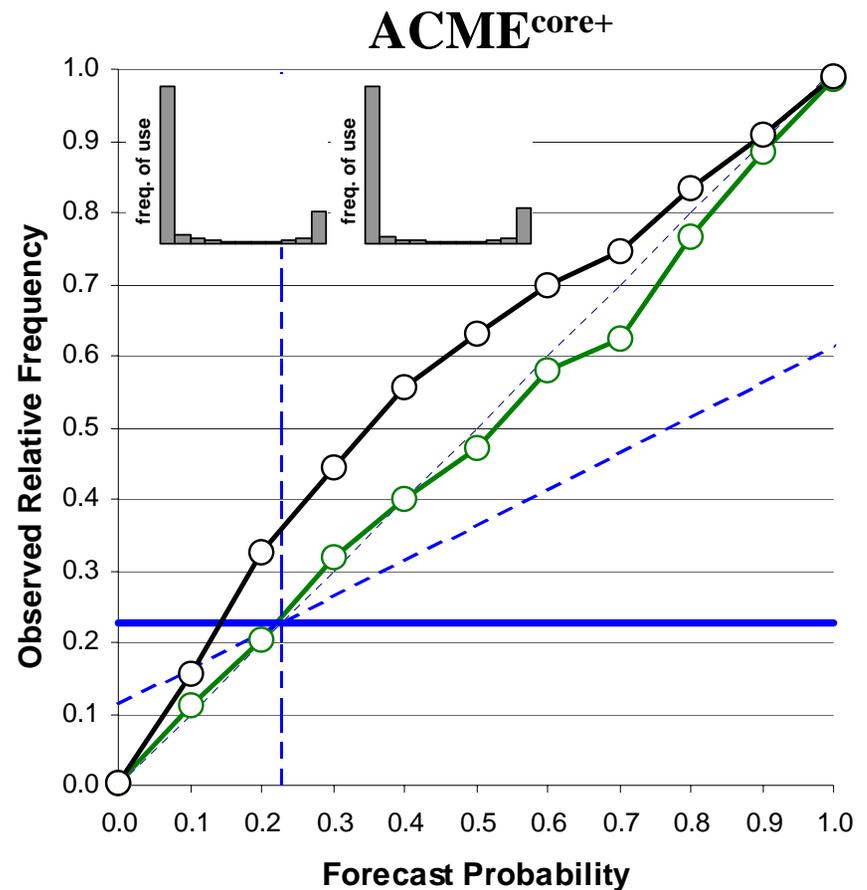
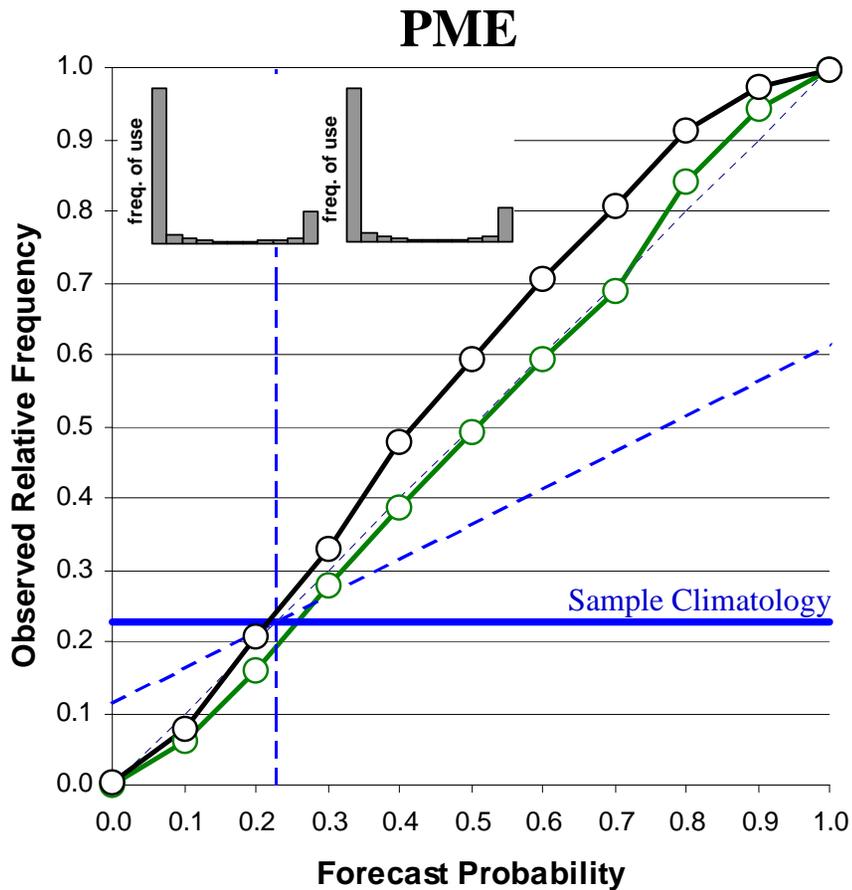
**\*Bias-corrected MSLP @ 36h**



EF Mean's MSE adjusted by  $n/n+1$  to account for small sample size

# Reliability Diagram Comparison

$P(MSLP < 1001\text{mb}) @ 36\text{h}$

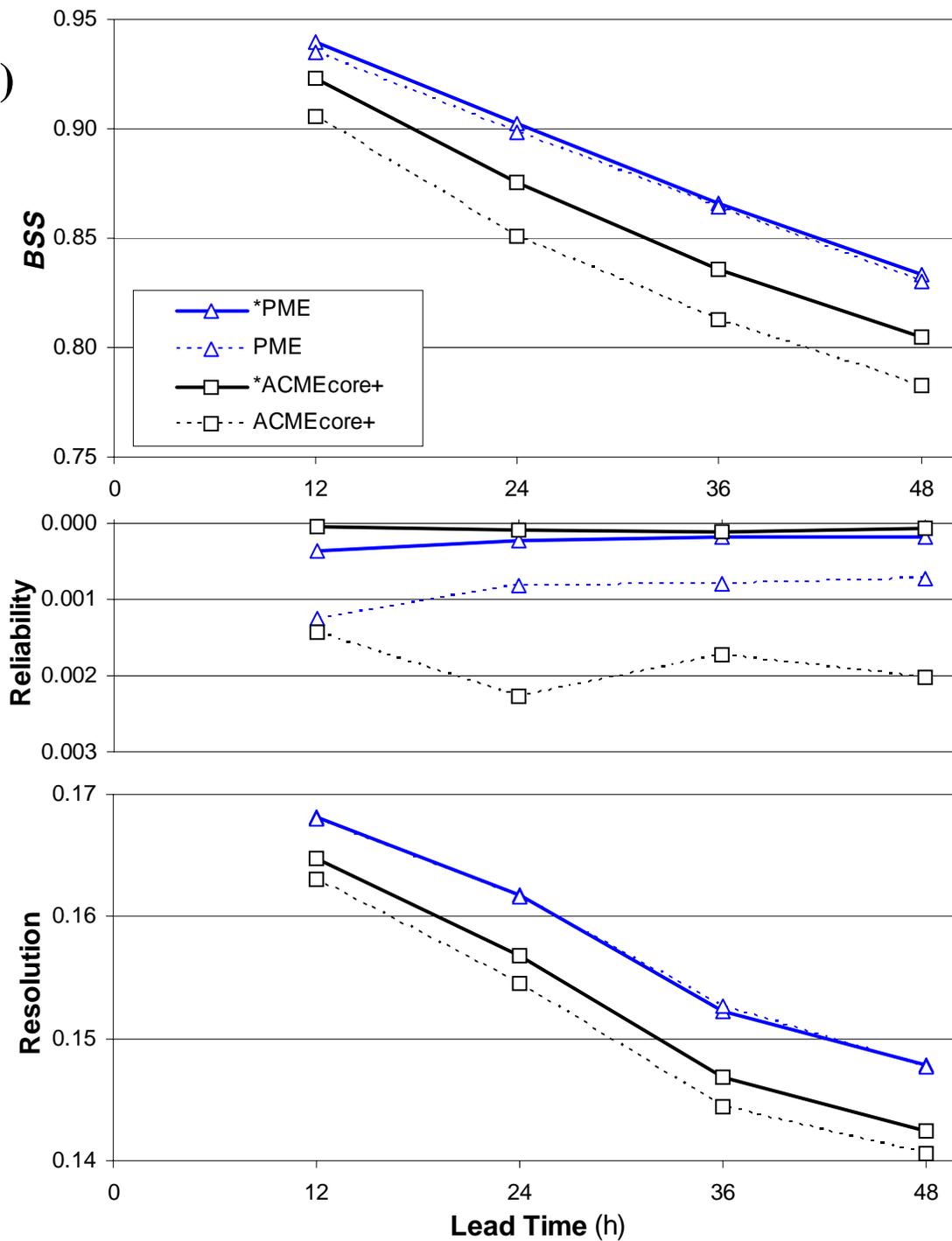


	Resolution	Reliability	Uncertainty	<i>BSS</i>
○—○ Biased	0.1526	0.0008	0.1756	0.8641
○—○ Corrected	0.1522	0.0002	0.1756	0.8655

	Resolution	Reliability	Uncertainty	<i>BSS</i>
○—○ Biased	0.1443	0.0013	0.1756	0.8138
○—○ Corrected	0.1465	0.0002	0.1756	0.8330

# Skill for $P(\text{MSLP} < 1001 \text{ mb})$

$$BSS = \frac{res - rel}{unc}$$



# How Well Does an EF Capture Truth?

- **Missing Rate (MR)** – The percentage of verifications not “encompassed” by the ensemble members

$$MR = \left( \frac{N_1 + N_{n+1}}{M} \right) 100$$

$N_1$  : Number of verifications in rank #1  
of a VRH

$N_{n+1}$  : Number of verifications in rank # $n+1$   
where  $n$  is the number of members

$M$  : Number of verifications

or **Missing Rate Error (MRE)**:

$$MRE = MR - \left( \frac{2}{n+1} \right) 100$$

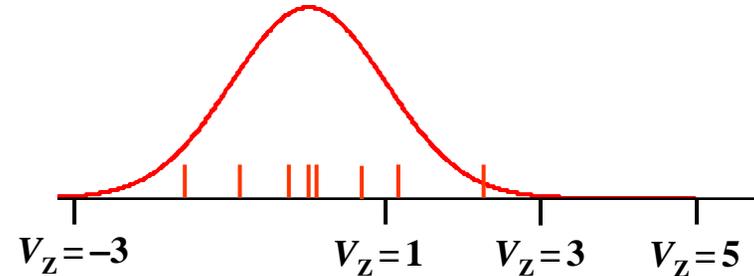
- **Standardized Verification ( $V_Z$ )** – The deviation of the verification with respect to the standard deviation of the ensemble members

$$V_Z = \frac{V - \bar{e}}{s}$$

$\bar{e}$  : Ensemble mean

$V$  : Verification value

$s$  : Ensemble standard deviation



- **Verification Outlier Percentage (VOP)** – The percentage of verifications not “portrayed” by the ensemble members (i.e., how often verification is an outlier)

$$VOP = \frac{100}{M} \sum_{m=1}^M \begin{cases} 0: & 3s_m \geq |V_m - \bar{e}_m| \\ 1: & 3s_m < |V_m - \bar{e}_m| \end{cases}$$

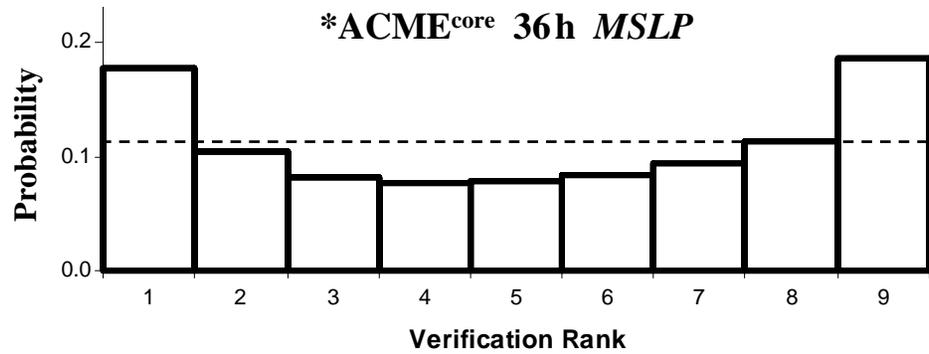
$\bar{e}_m$  : Ensemble mean at point  $m$

$V_m$  : Verification value at point  $m$

$s_m$  : Ensemble standard deviation at point  $m$

$E\{VOP\} \cong 0.3\%$   
for large  $n$  and  
normal PDF

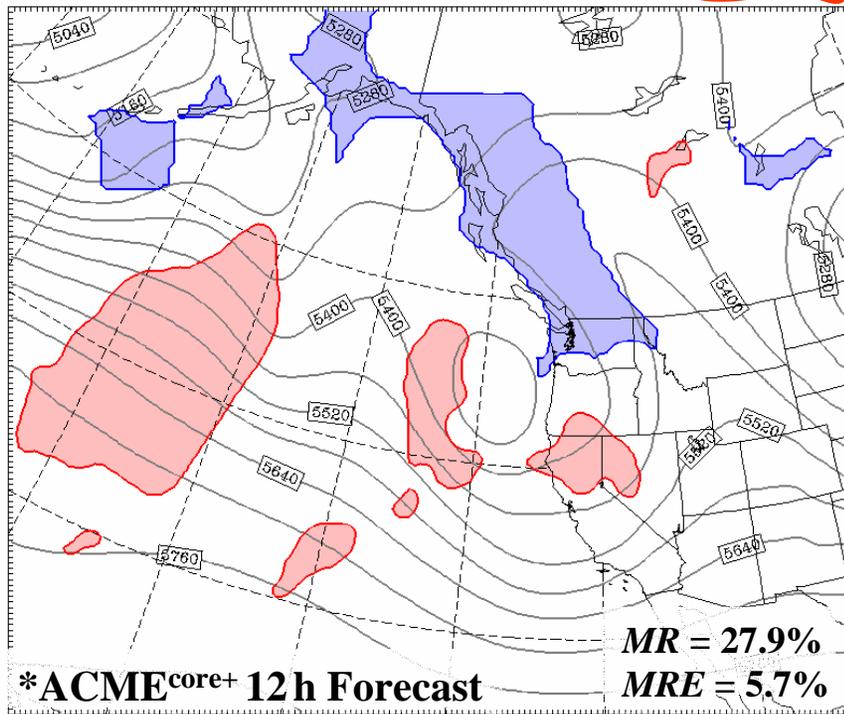
# Sample Applications



**MR = 36.6%**  
**MRE = 14.4%**  
**VOP = 9.0%**

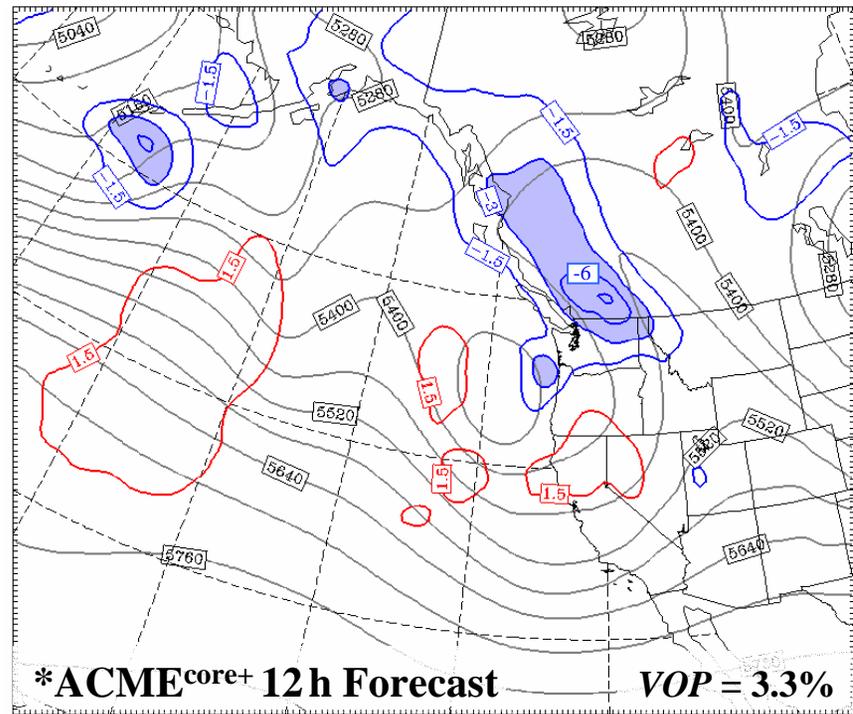
**Z<sub>500</sub> EF Mean  
 & Verification Extreme Ranks**

rank 1 (blue circle)  
 rank 9 (red circle)



**Z<sub>500</sub> EF Mean & V<sub>z</sub>**

V<sub>z</sub> < -3 (blue circle)  
 V<sub>z</sub> > 3 (red circle)



Case Study Initialized at 00Z, 20 Dec 2002

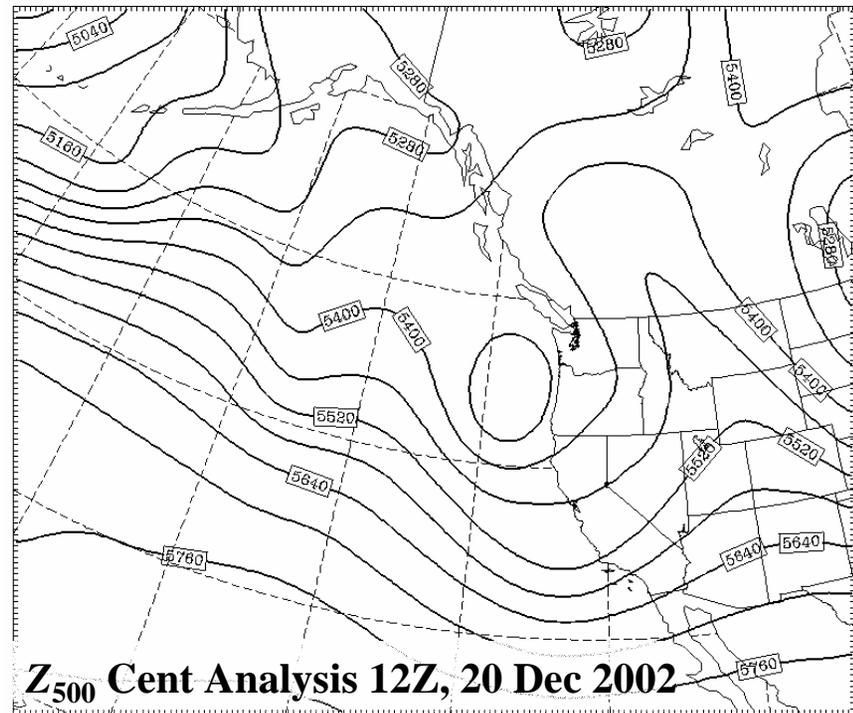
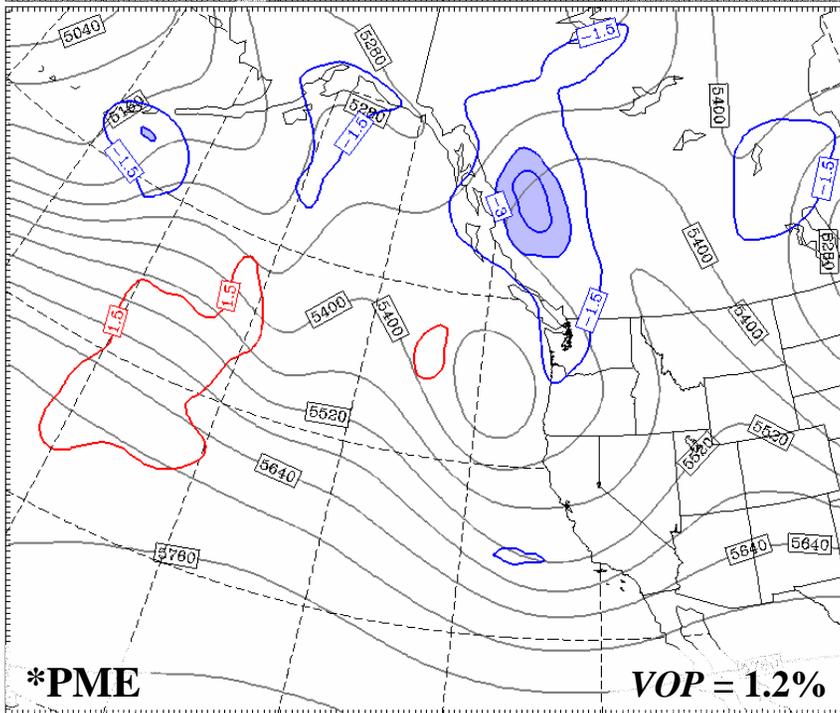
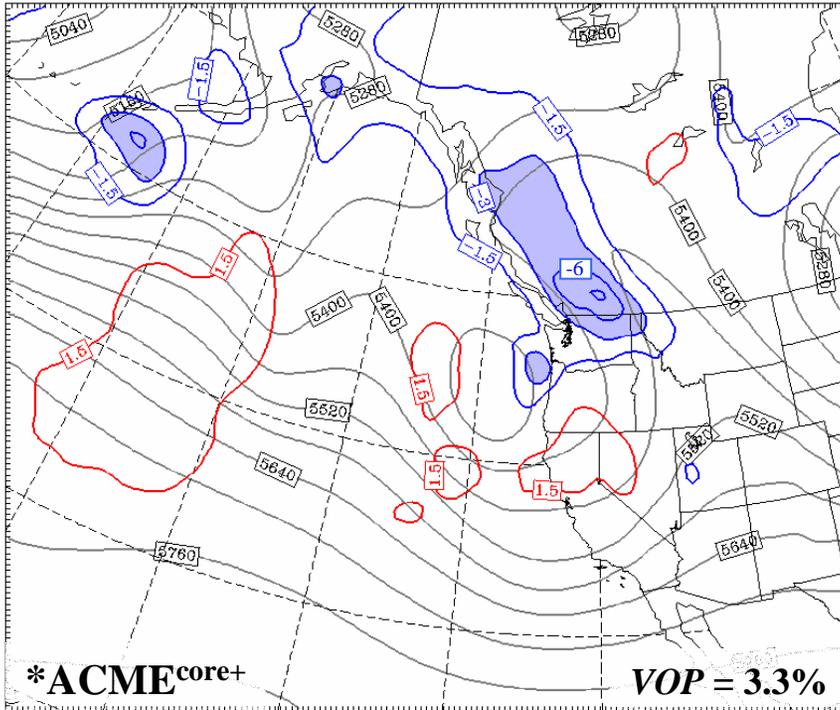
**12h Forecast**

**Z<sub>500</sub> EF Mean and Standardized Verification, V<sub>z</sub>**

**V<sub>z</sub> < -3**

**V<sub>z</sub> > 3**

$$V_z = \frac{V - \bar{e}}{s}$$



Case Study Initialized at 00Z, 20 Dec 2002

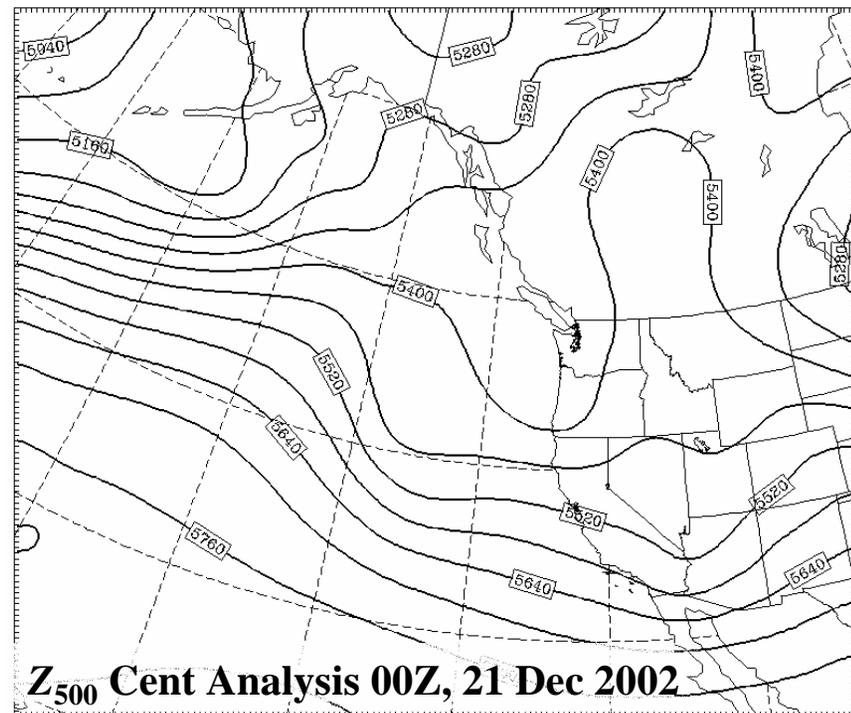
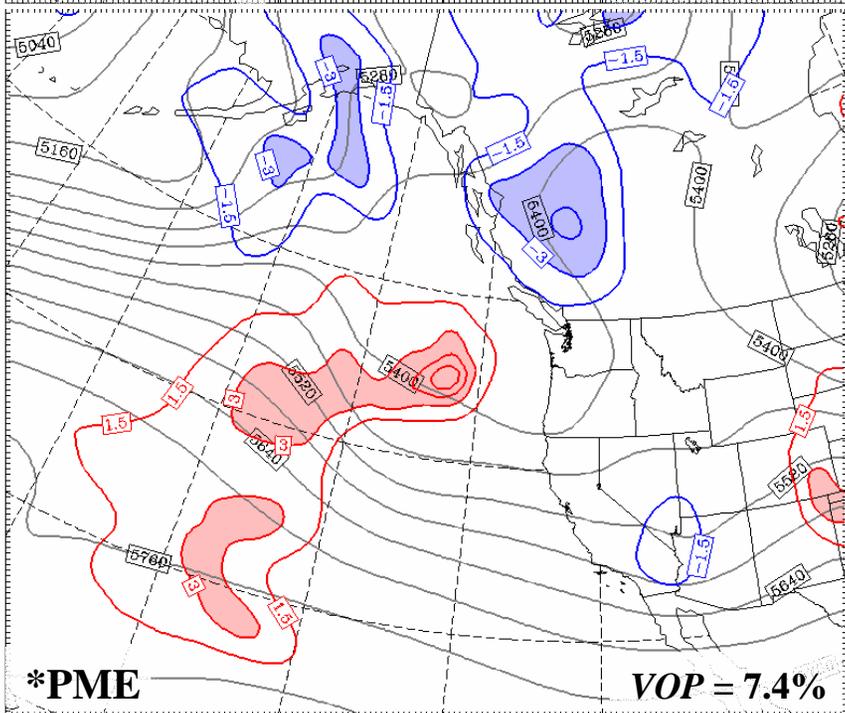
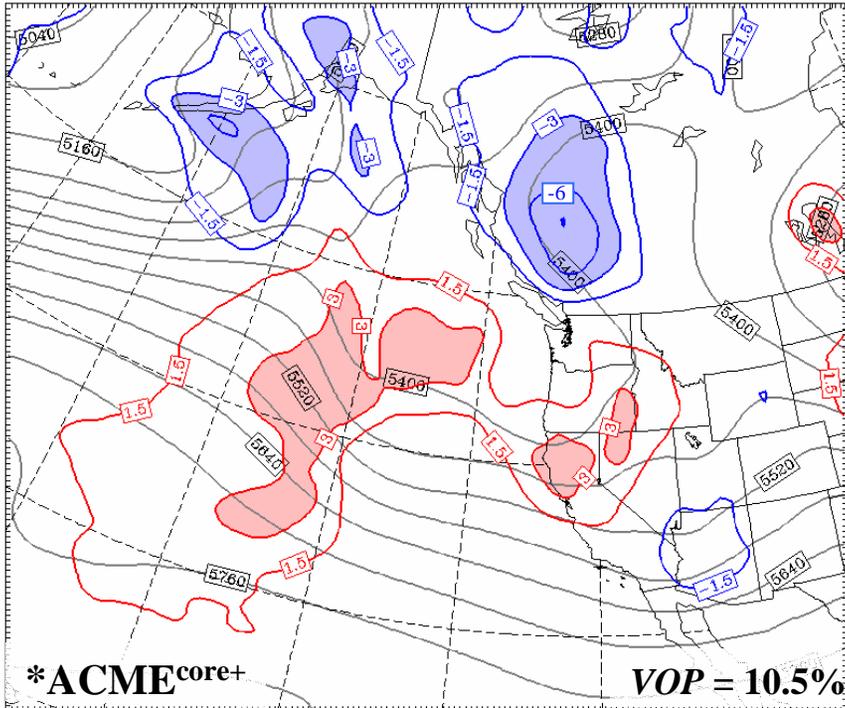
**24h Forecast**

**Z<sub>500</sub> EF Mean and Standardized Verification, V<sub>z</sub>**

**V<sub>z</sub> < -3**

**V<sub>z</sub> > 3**

$$V_z = \frac{V - \bar{e}}{s}$$



Case Study Initialized at 00Z, 20 Dec 2002

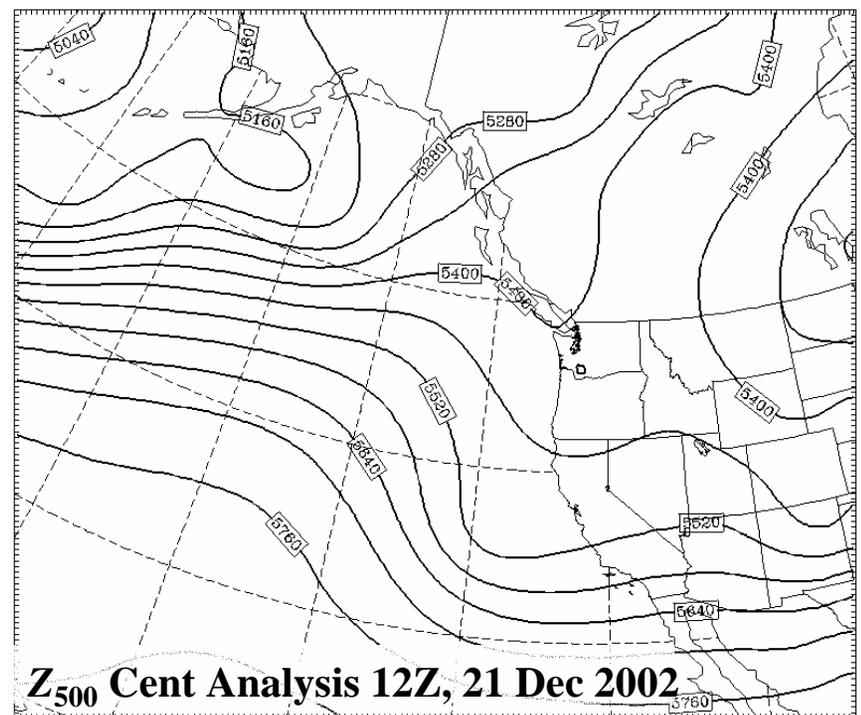
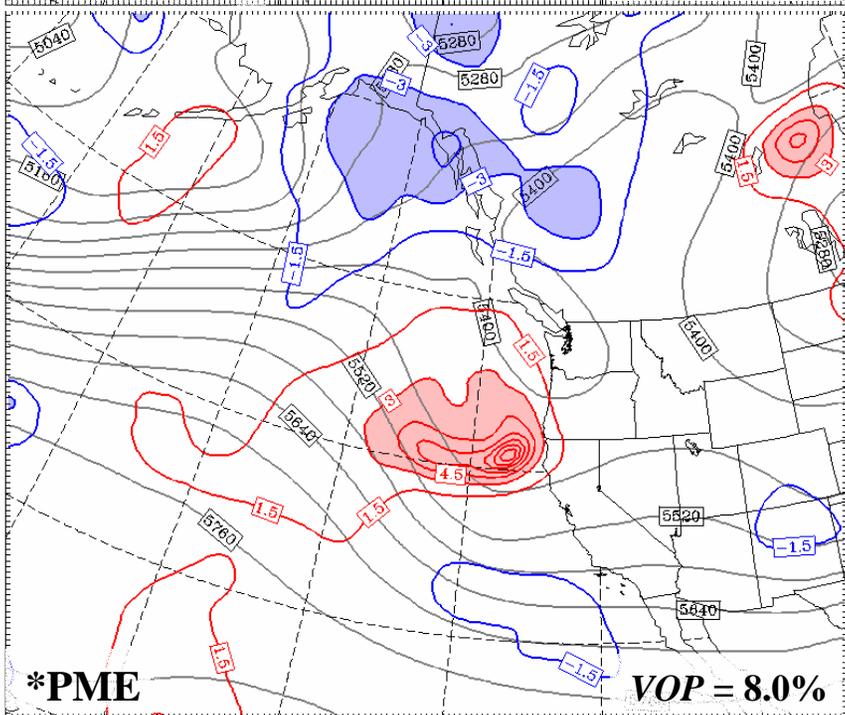
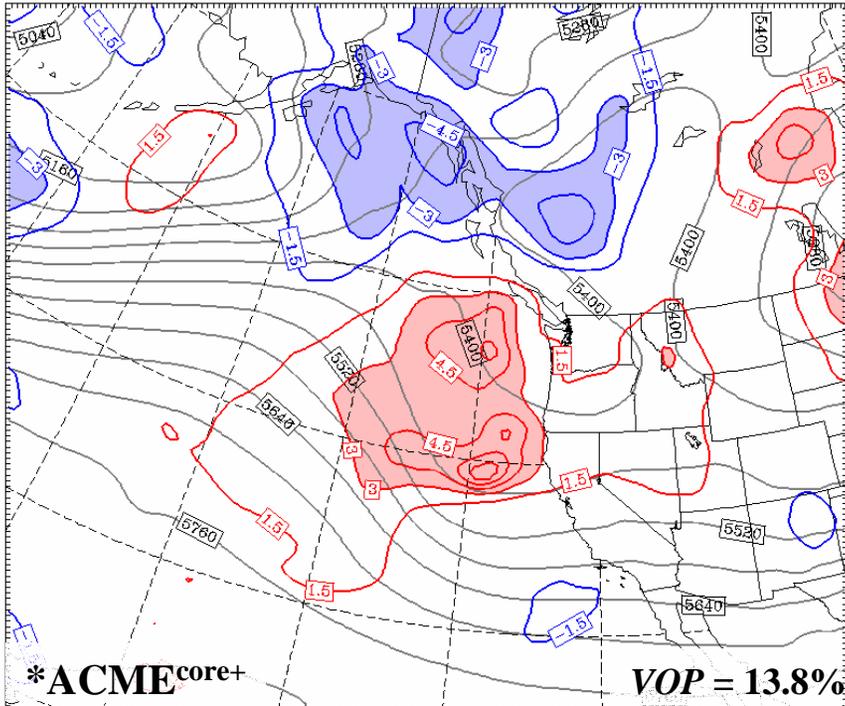
**36h Forecast**

**Z<sub>500</sub> EF Mean and Standardized Verification, V<sub>z</sub>**

**V<sub>z</sub> < -3**

**V<sub>z</sub> > 3**

$$V_z = \frac{V - \bar{e}}{s}$$



Case Study Initialized at 00Z, 20 Dec 2002

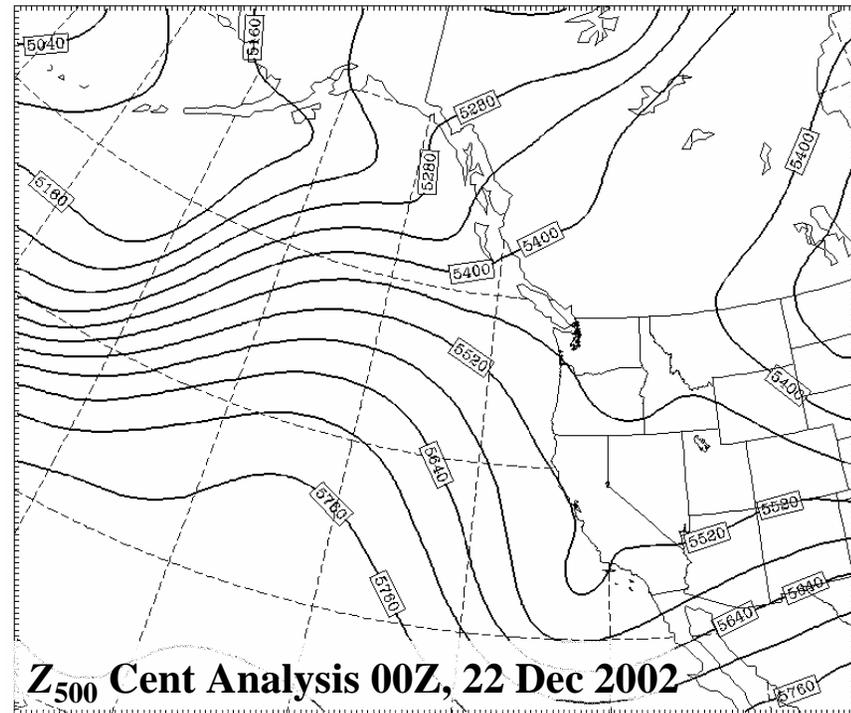
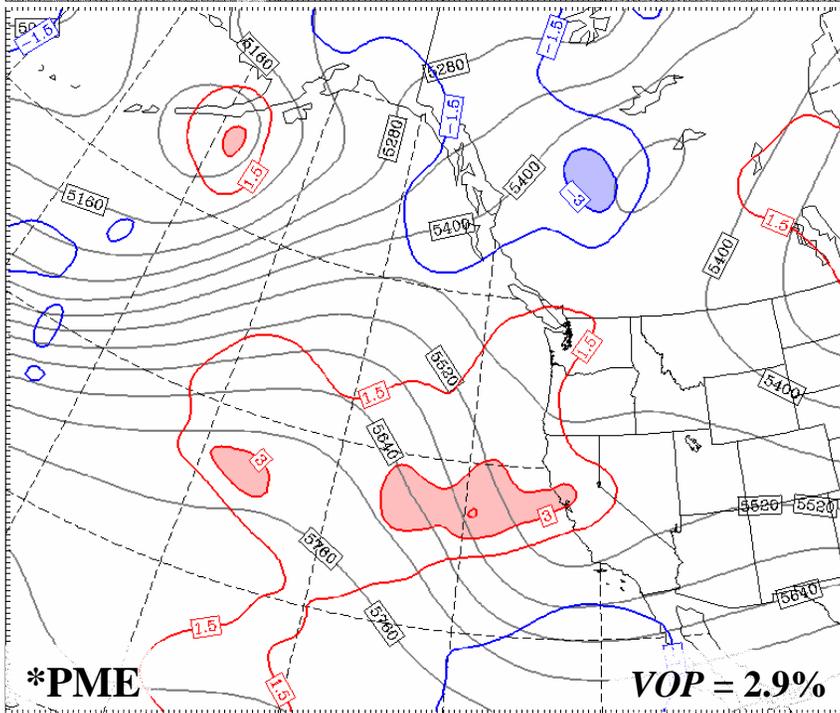
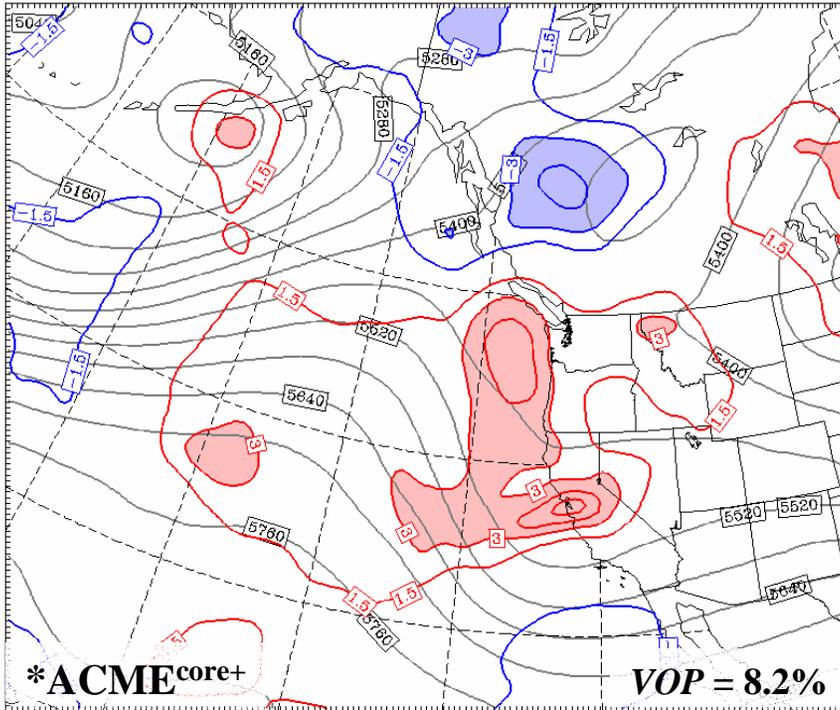
**48h Forecast**

**Z<sub>500</sub> EF Mean and Standardized Verification, V<sub>z</sub>**

**V<sub>z</sub> < -3**

**V<sub>z</sub> > 3**

$$V_z = \frac{V - \bar{e}}{s}$$



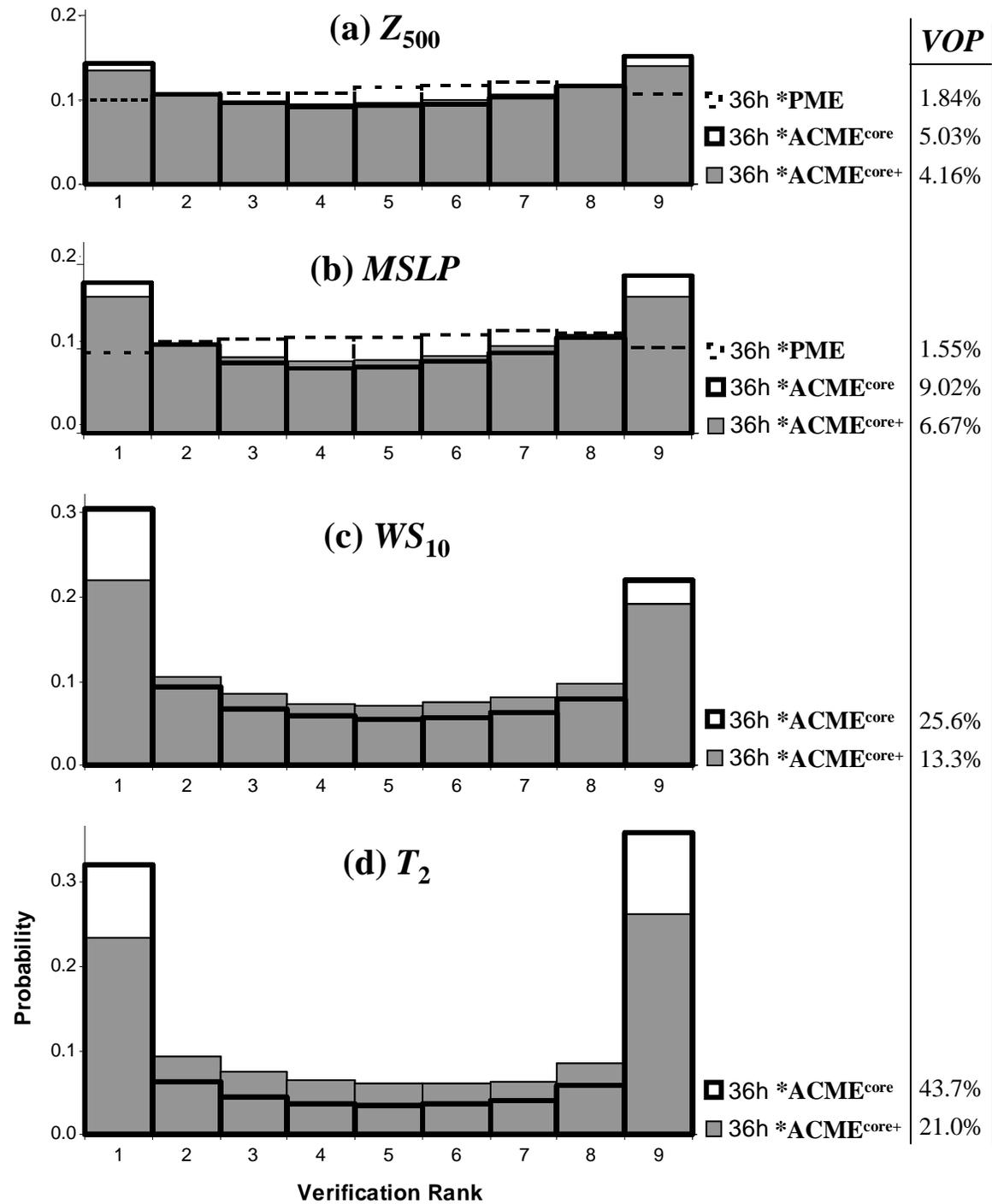
**Value of  
Model Diversity  
For a Mesoscale SREF**

**ACME<sup>core</sup>  
Vs.  
ACME<sup>core+</sup>**



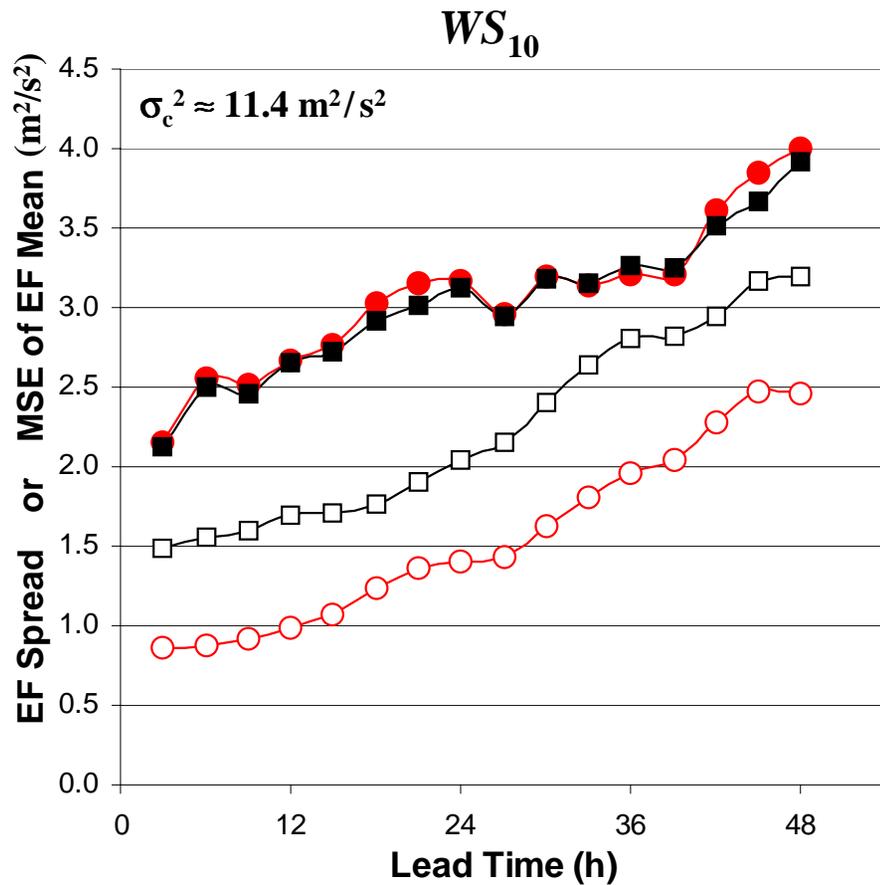
  
**Synoptic  
Parameter**  
 (Errors Depend on  
IC Uncertainty)

  
**Surface/Mesoscale  
Parameter**  
 (Errors Depend on  
Model Uncertainty)

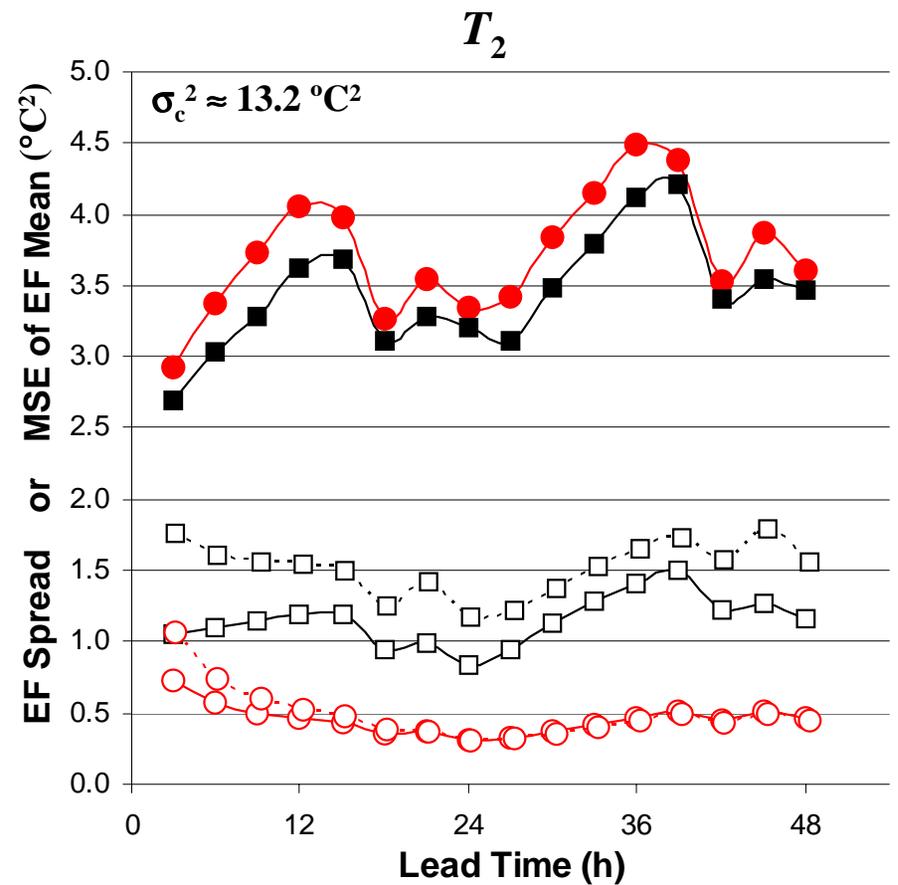


# Ensemble Dispersion

\*Bias Corrected

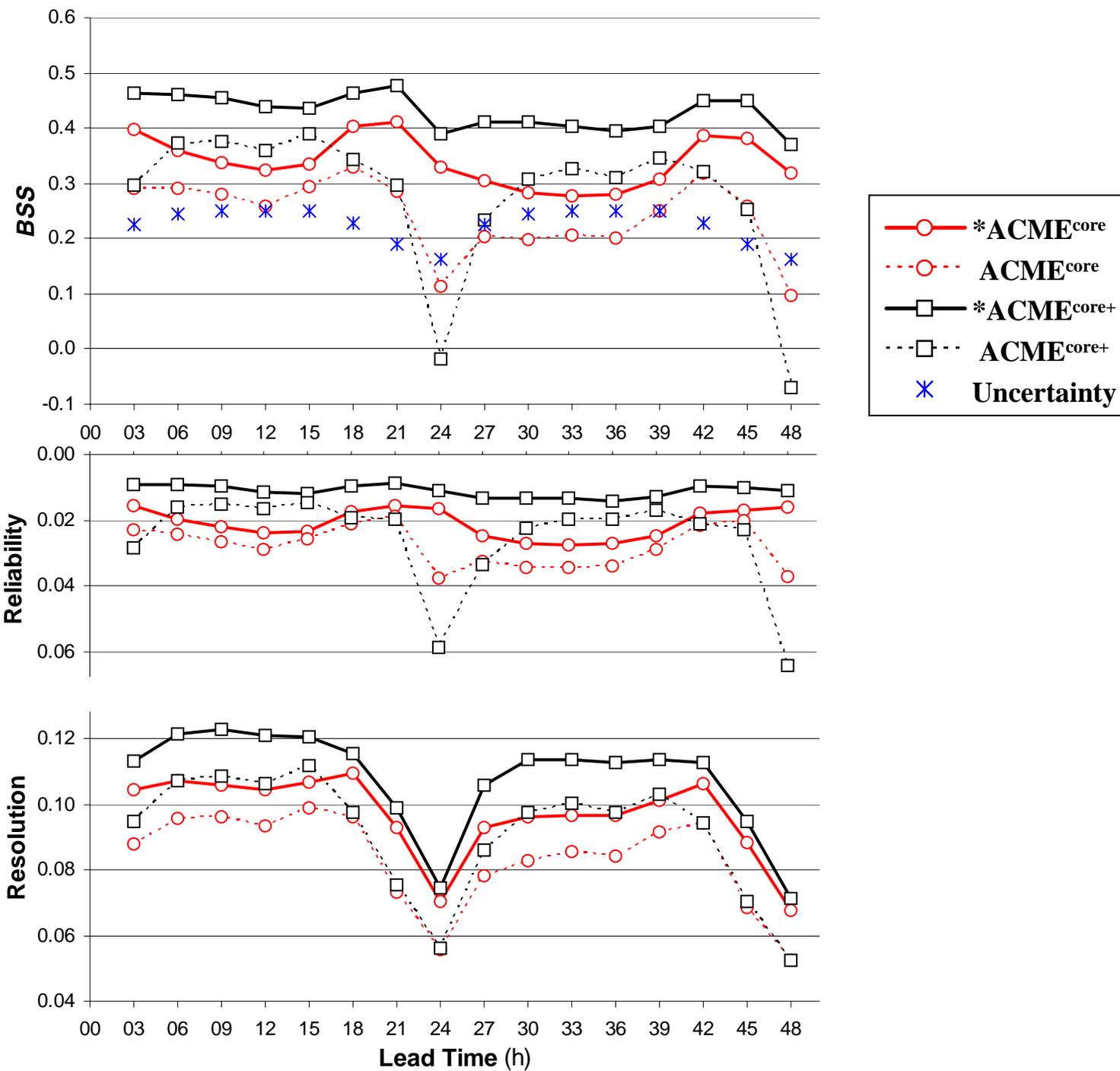


- \*ACME<sup>core</sup> Spread
- MSE of \*ACME<sup>core</sup> Mean
- \*ACME<sup>core+</sup> Spread
- MSE of \*ACME<sup>core+</sup> Mean



- ACME<sup>core</sup> Spread (Uncorrected)
- ACME<sup>core+</sup> Spread (Uncorrected)

# Skill for $P(T_2 < 0^\circ\text{C})$



## Findings (Old & New)

- 1) Bias correction improves EF skill (Richardson, 2001; Atger, 2003)
  - a) **Particularly important for mesoscale SREF in which model biases are often large**
  - b) Reliability is improved by correctly shifting the forecast PDF location
  - c) **Resolution is improved through increased sharpness of the forecast PDF**
  - d) **Allows for fair and accurate analysis (i.e., consider statistical consistency)**

## Findings (Old & New)

- 2) Inclusion of model diversity in an EF system is critical for complete representation of uncertainty (Houtekamer et al., 1996; Stensrud et al., 2000; Mylne et al., 2002 )
  - a) Increases the spread of an underdispersive EF to improve statistical consistency
  - b) **Degree of importance depends on parameter, scale, event of interest, and weather regime**
  - c) Unequal skill among ensemble members is not problematic (Mylne, 2002)
  - d) **Both reliability and resolution are improved (even with increased spread) through better estimation of the forecast PDF**
  - e) A PME or multimodel system is good at capturing model uncertainty (Evans et al., 2000; Ziehmann, 2000; Ebert, 2001; Richardson, 2001; Wandishin et al., 2001; Mylne, 2002)
    - **Does a more thorough job than a perturbed-model system**
    - **Slight overdispersion in a small EF may be advantageous**

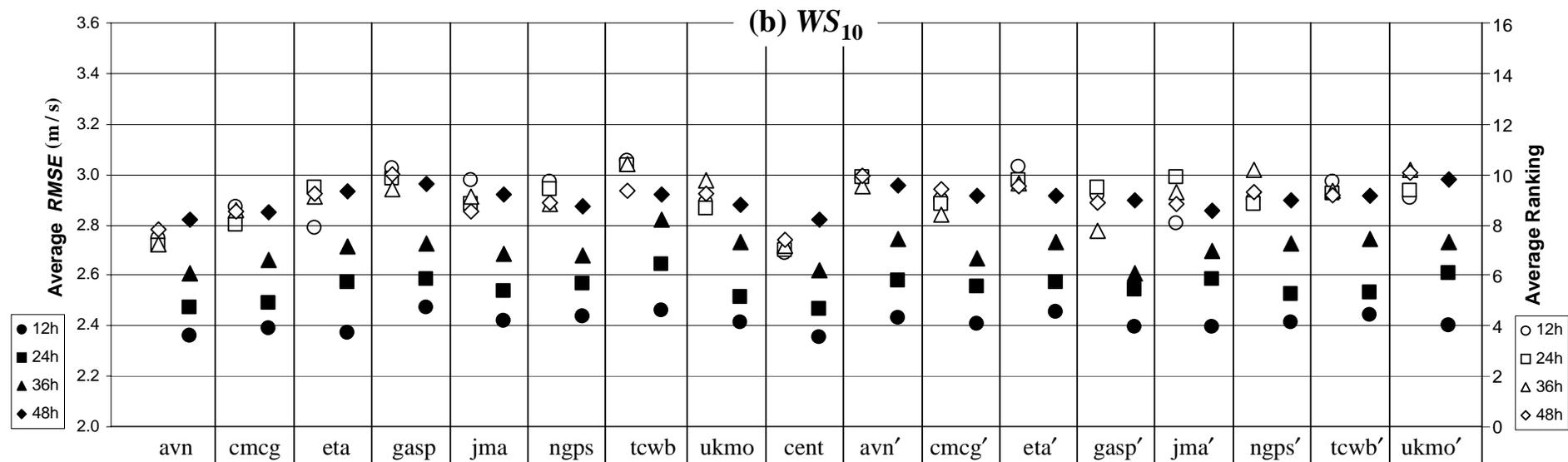
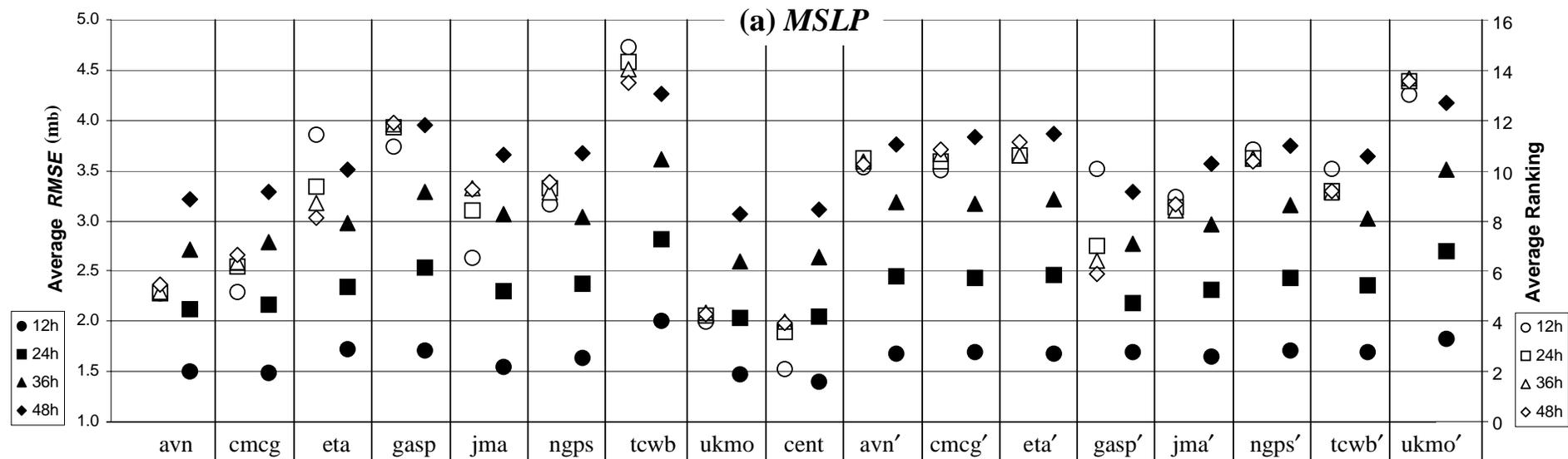
## References

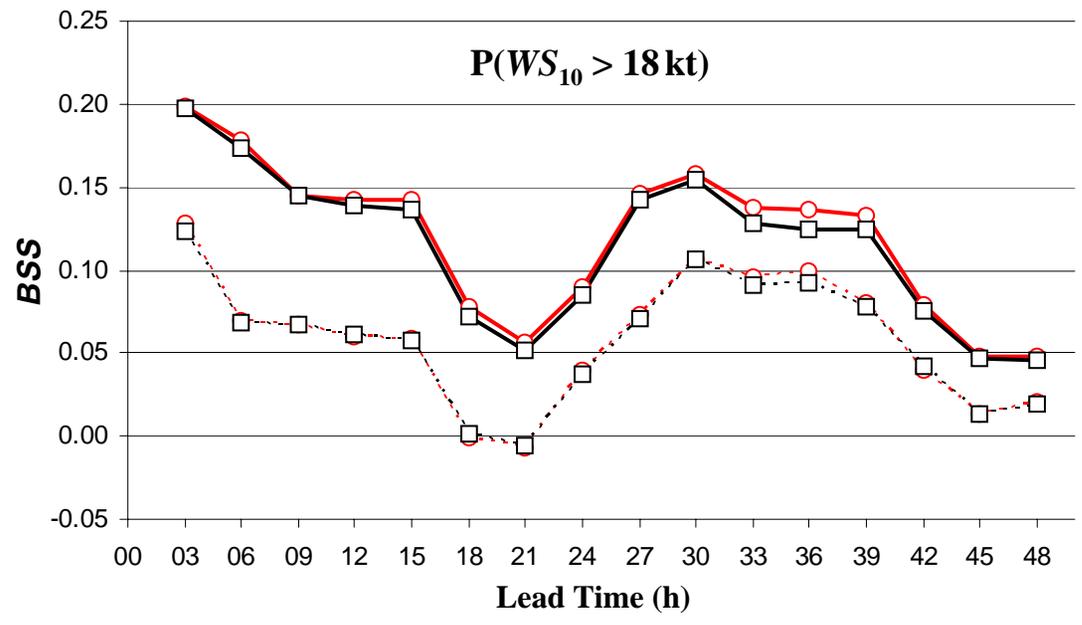
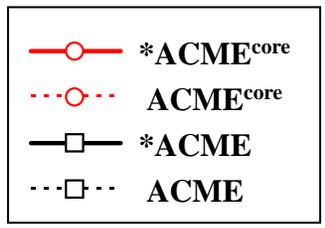
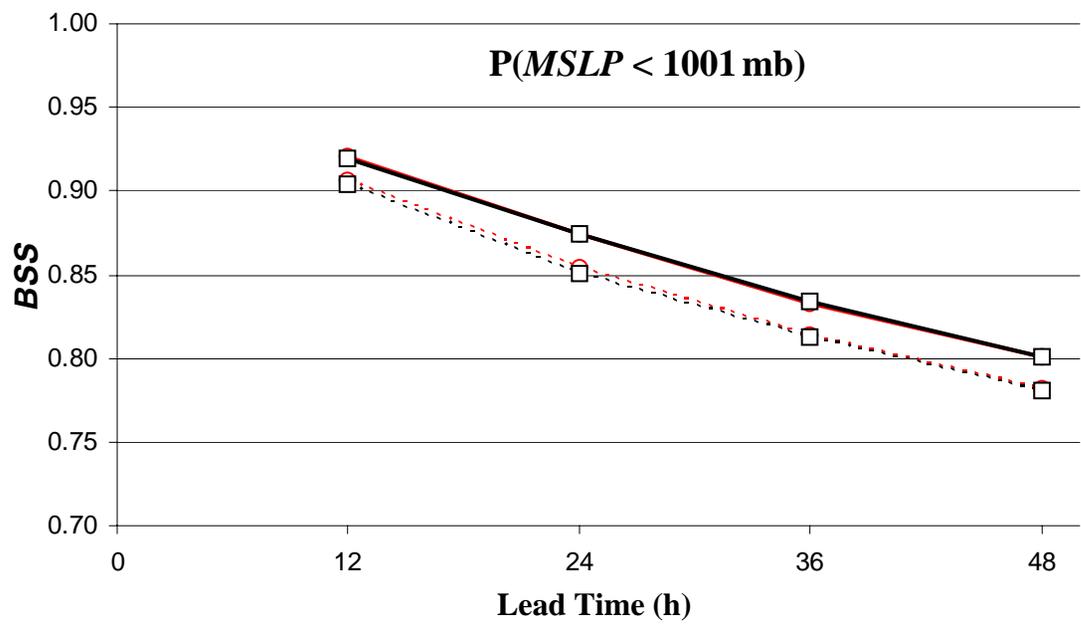
- Atger, F, 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Wea. Rev.*, **121**, 1509–1523.
- Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.
- Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104–3127.
- Houtekamer, P.L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196.
- Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multimodel multianalysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.*, **128**, 361–384.
- Richardson, D. S., 2001a: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.*, **127**, 1847–1864.
- Wandishin, M. S., S L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.
- Ziehmann, C., 2000: Comparison of a single-model EPS with a multimodel ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299

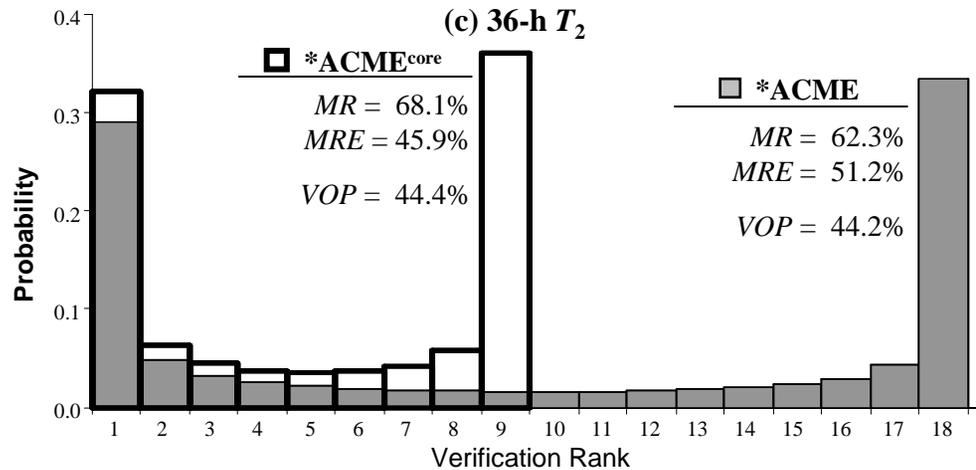
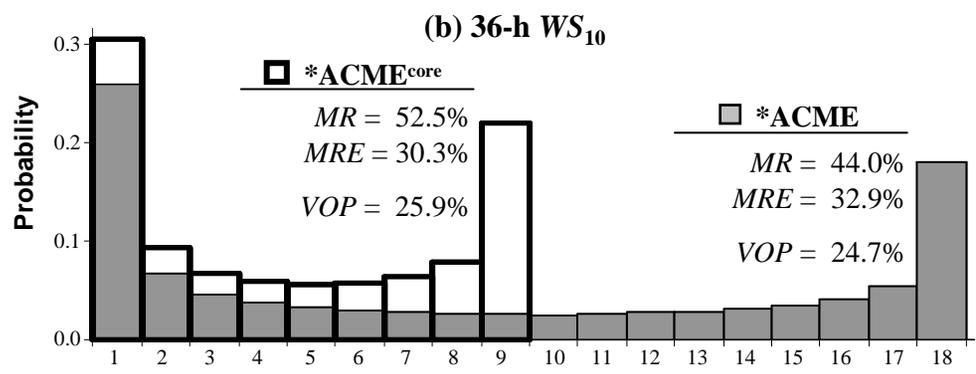
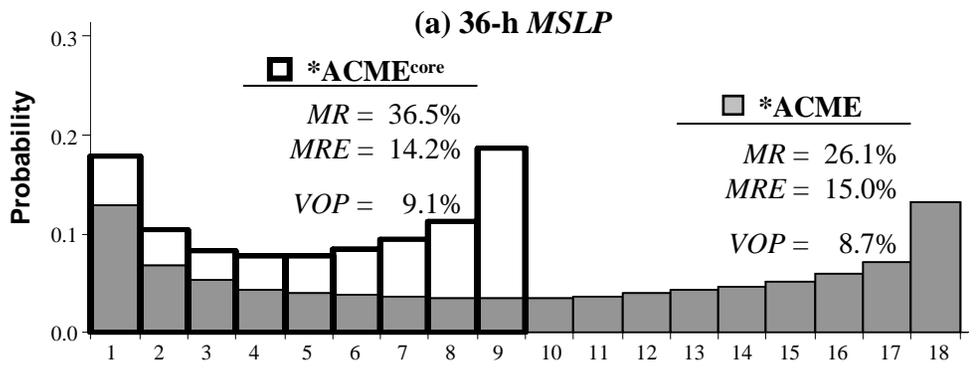
*Backup Slides*

**Success and Failure  
of  
ACME**

**ACME<sup>core</sup>  
Vs.  
ACME**







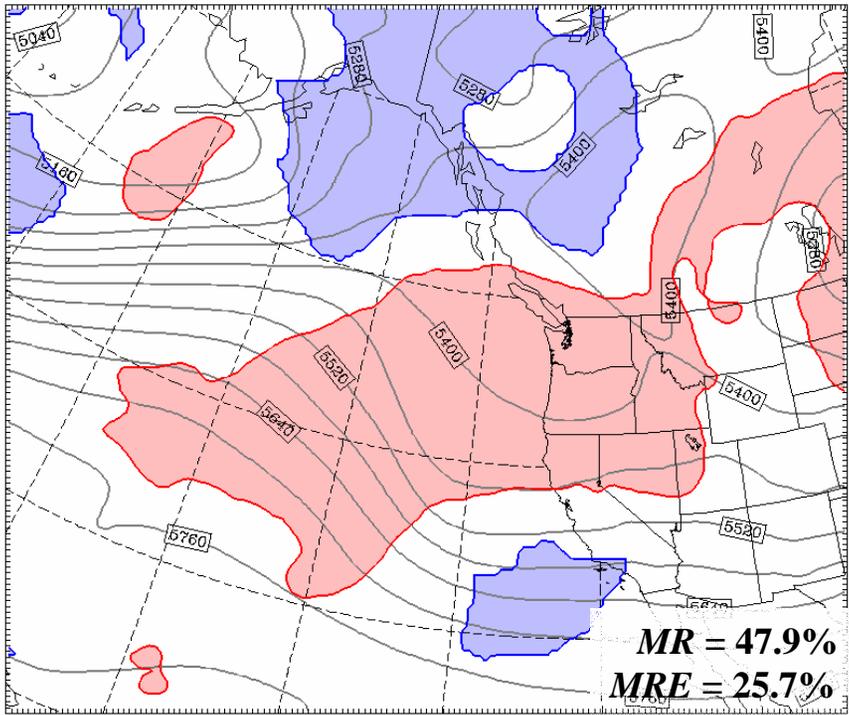
# Case Study Initialized at 00Z, 20 Dec 2002

## 36h Forecast

**\*ACME<sub>core</sub>**

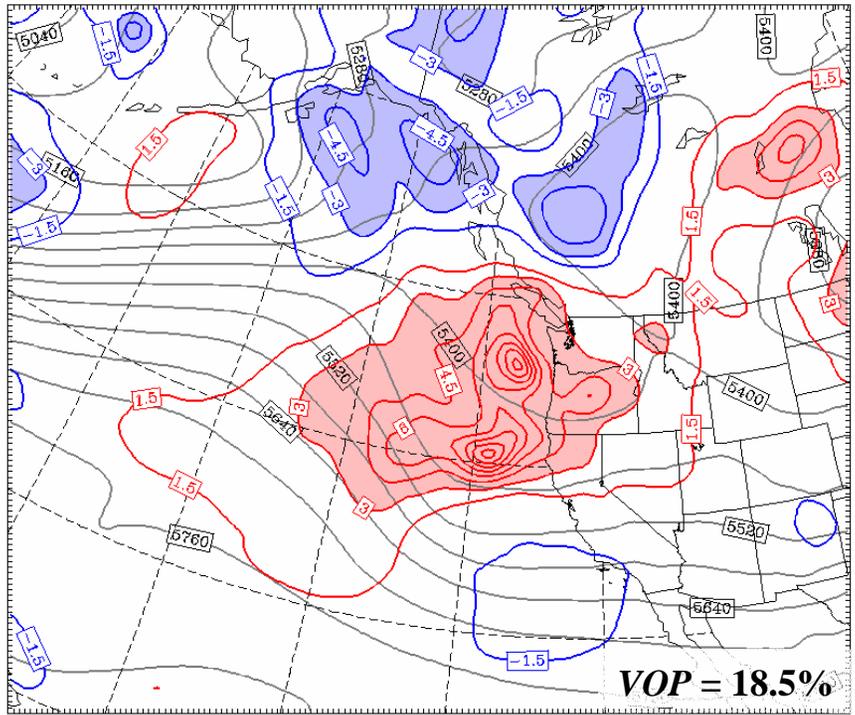
### Z<sub>500</sub> EF Mean and Verification Extreme Ranks

rank 1      rank 9



### Z<sub>500</sub> EF Mean and V<sub>z</sub>

V<sub>z</sub> > 3      V<sub>z</sub> < -3



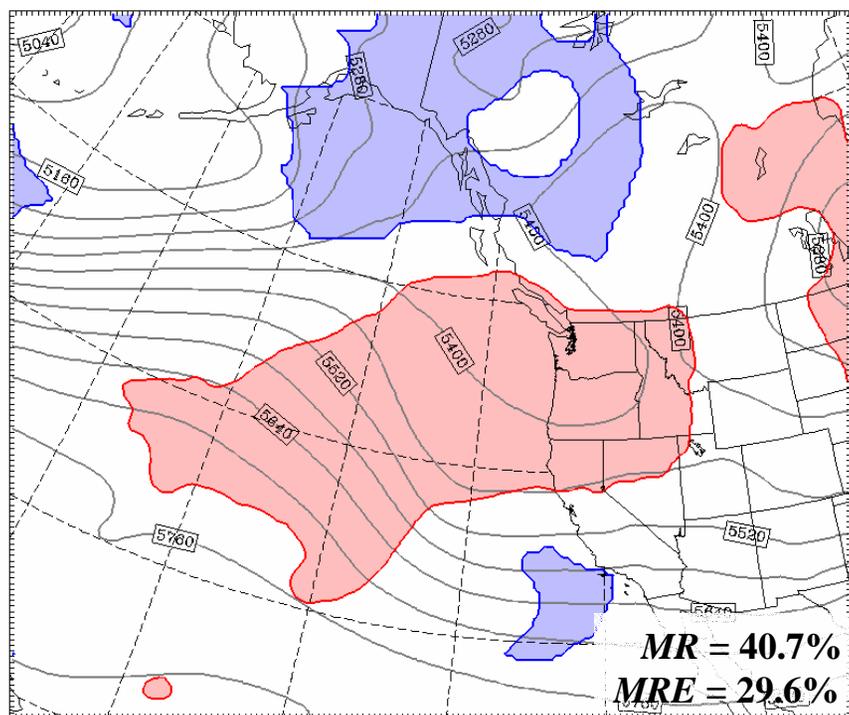
# Case Study Initialized at 00Z, 20 Dec 2002

## 36h Forecast

**\*ACME**

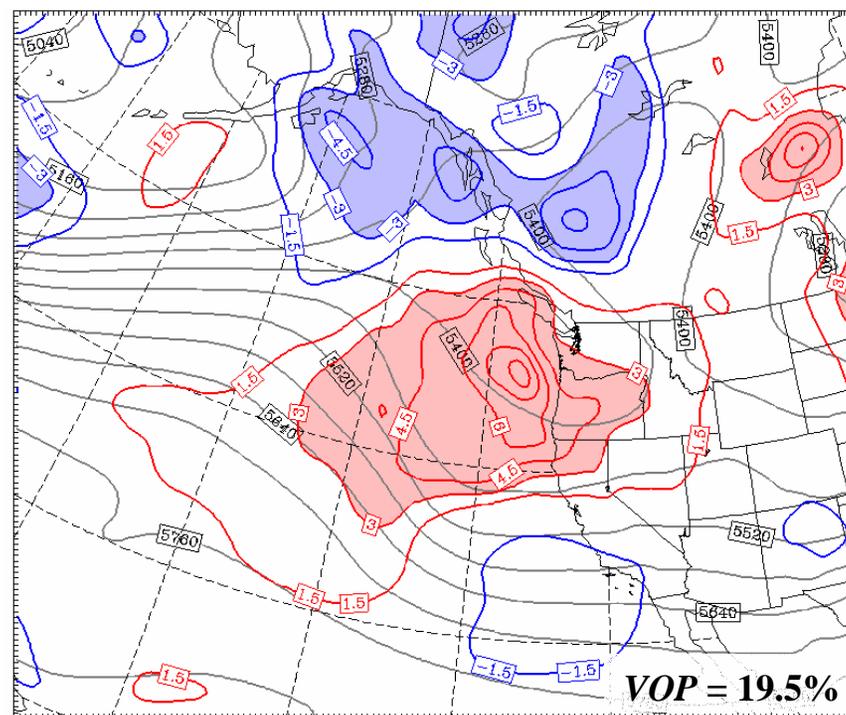
### Z<sub>500</sub> EF Mean and Verification Extreme Ranks

rank 1      rank 18



### Z<sub>500</sub> EF Mean and V<sub>z</sub>

V<sub>z</sub> > 3      V<sub>z</sub> < -3



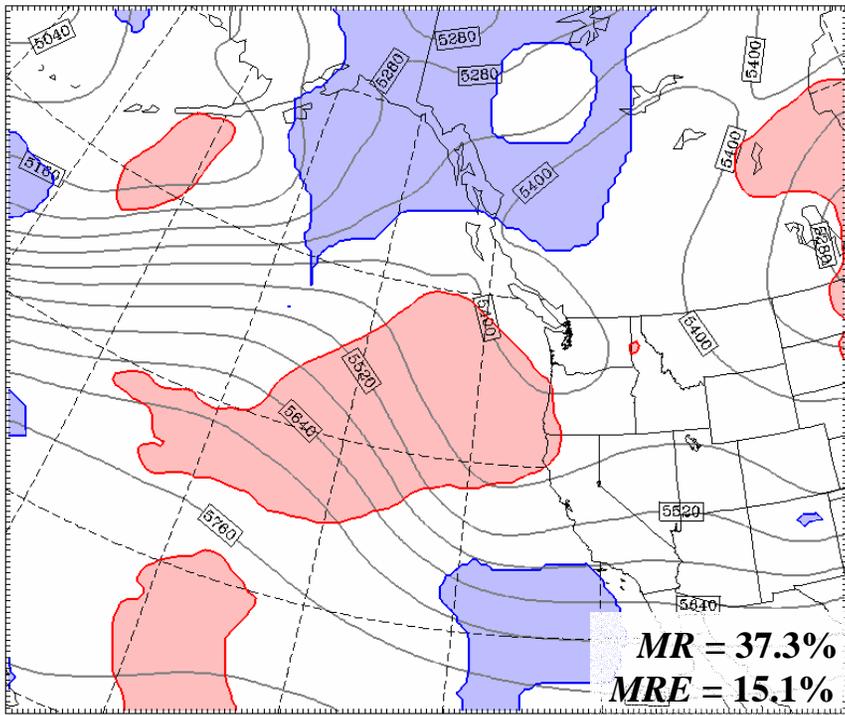
Case Study Initialized at 00Z, 20 Dec 2002

36h Forecast

**\*PME**

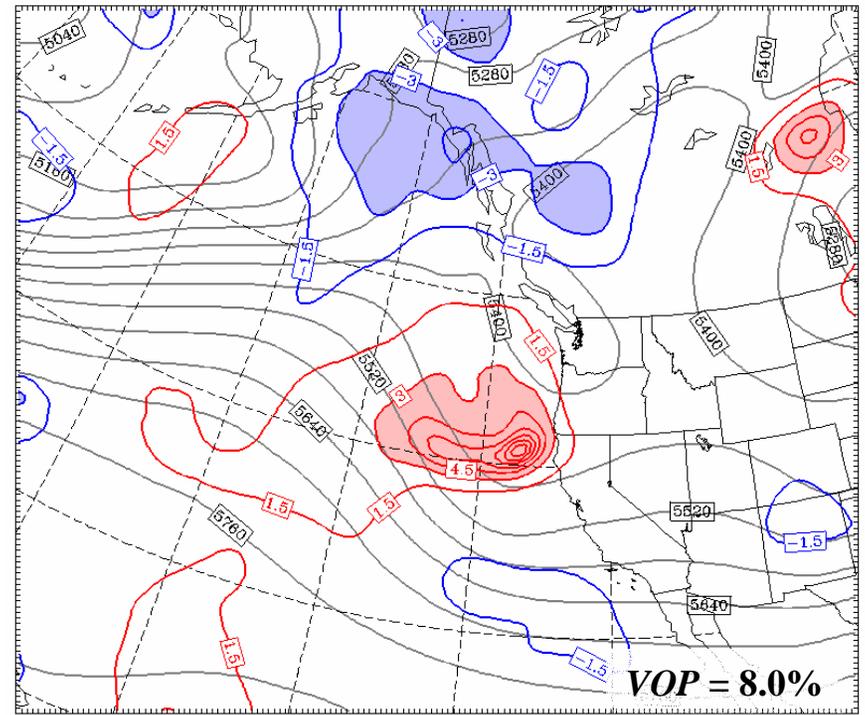
**Z<sub>500</sub> EF Mean and  
Verification Extreme Ranks**

rank 1      rank 9



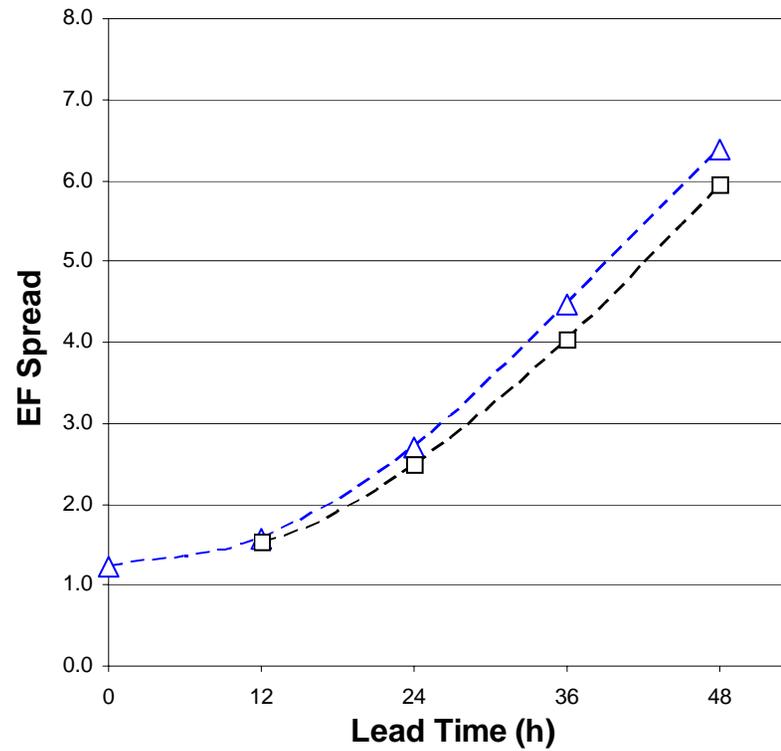
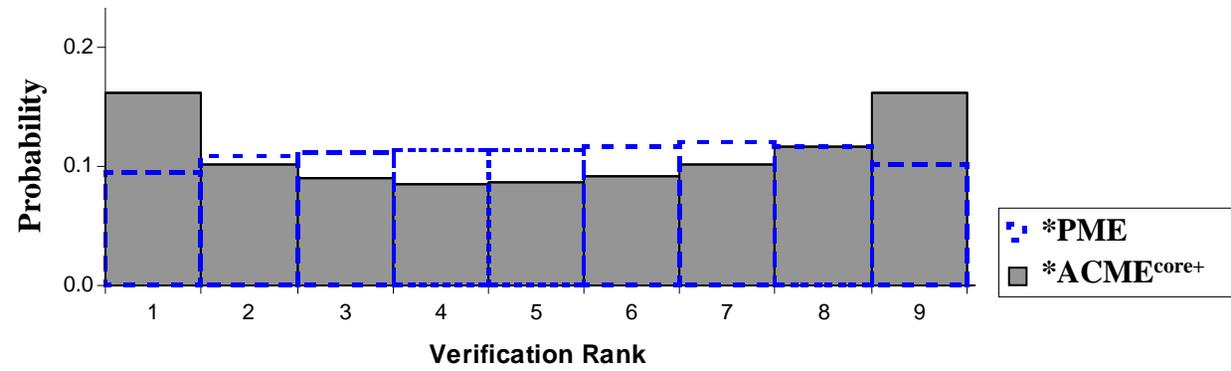
**Z<sub>500</sub> EF Mean and V<sub>z</sub>**

V<sub>z</sub> > 3      V<sub>z</sub> < -3

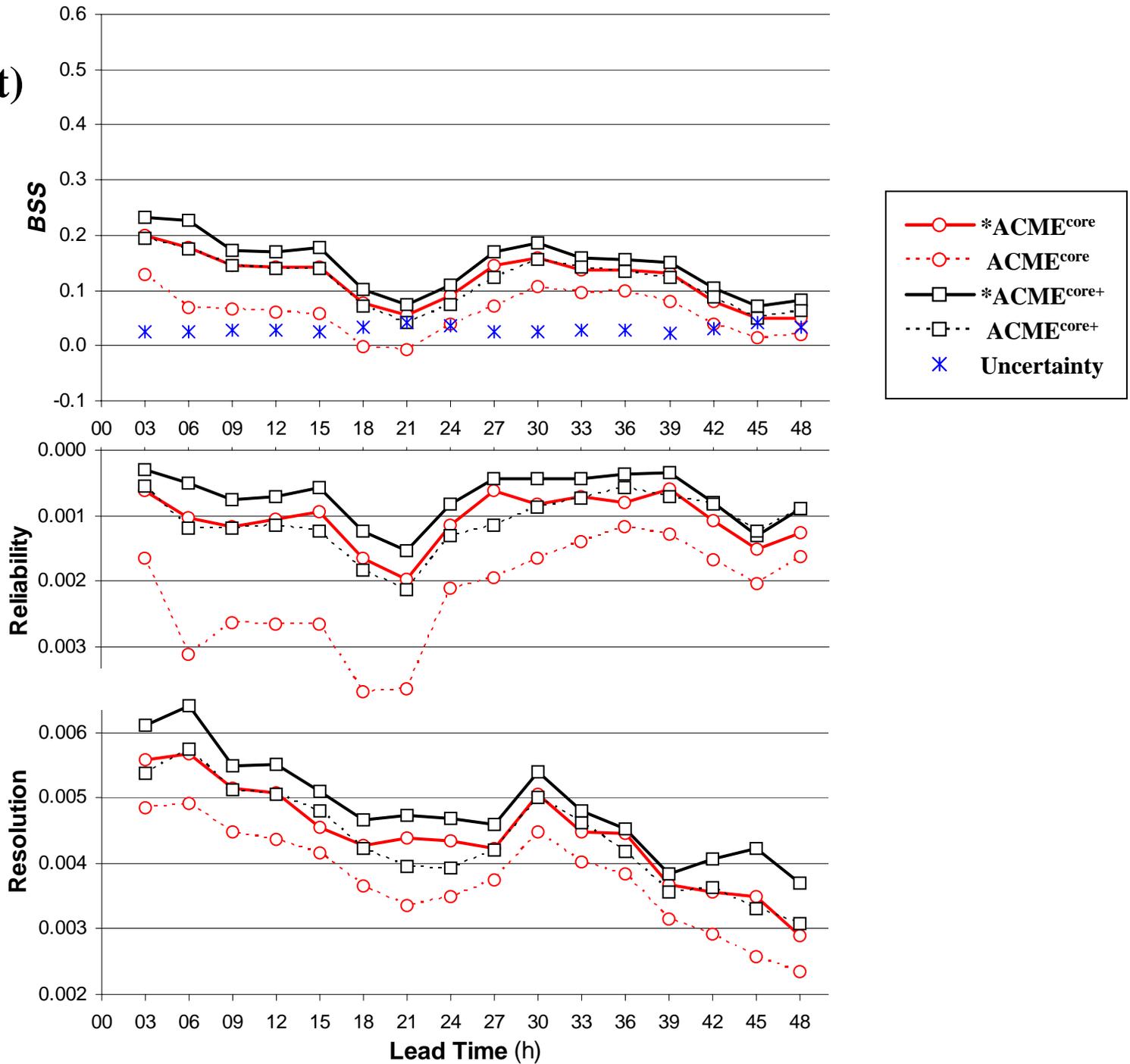


# Ensemble Dispersion

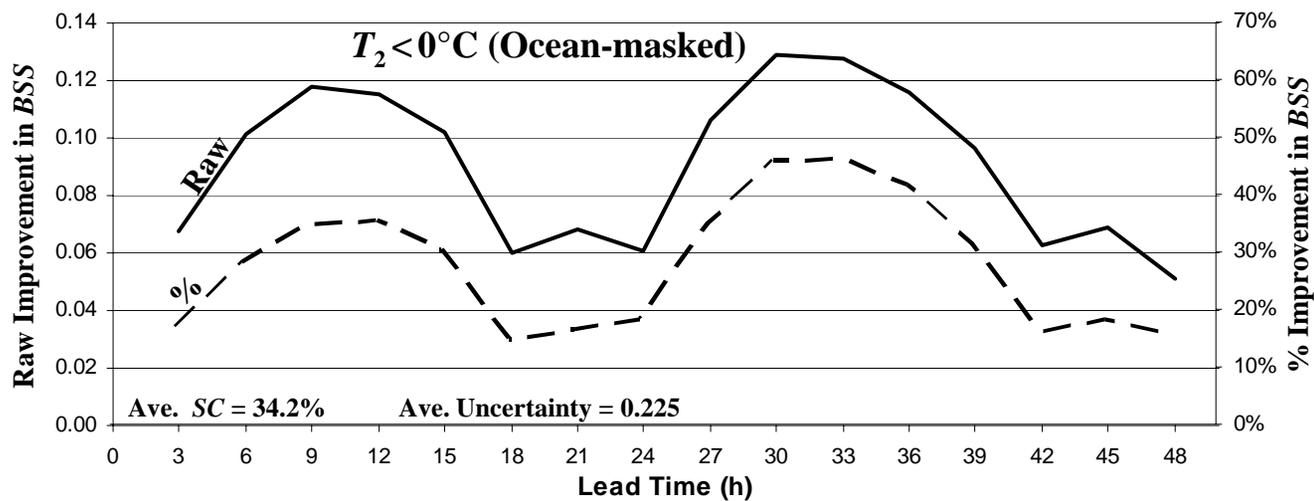
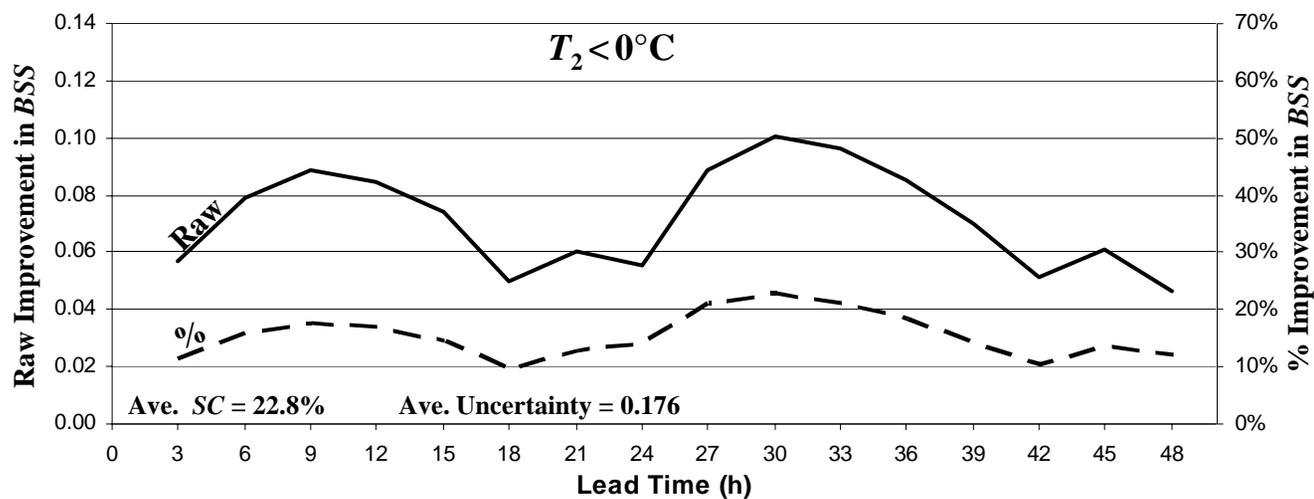
**\*Bias-corrected *MSLP* @ 36 h**



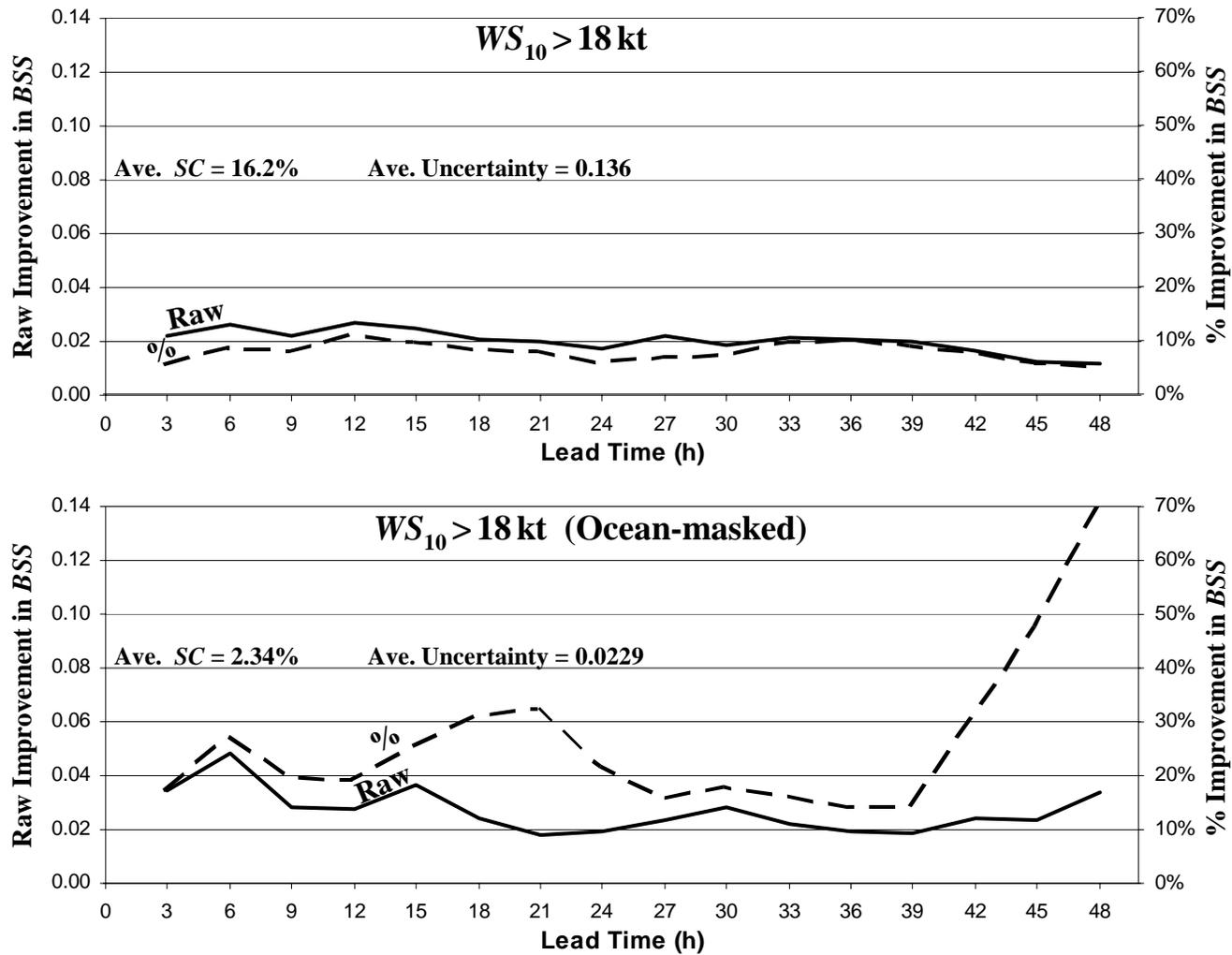
# Skill for $P(W_{S_{10}} > 18 \text{ kt})$



# BSS Improvement of ACME<sup>core+</sup> over ACME<sup>core</sup>



# BSS Improvement of ACME<sup>core+</sup> over ACME<sup>core</sup>



# Probabilistic Forecast Verification

## Continuous

### Brier Score

$$BS = \frac{1}{n} \sum_{i=1}^n (FP_i - OBS_i)^2$$

$n$ : number of data pairs

$FP_i$ : forecast probability {0.0...1.0}

$ORF_i$ : observation {0.0=yes, 1.0=no}

$BS = 1$  for perfect forecasts

$BS = 0$  for worst case forecasts

### Brier Skill Score

$$BSS = \frac{BS - BS_{\text{clim}}}{BS_{\text{perfect}} - BS_{\text{clim}}} = 1 - \frac{BS}{BS_{\text{clim}}}$$

$BSS = 1$  for perfect forecasts

$BSS < 0$  for forecasts worse than climo

## by Discrete Bins

### Decomposed Brier Score

$$BS = \underbrace{\frac{1}{n} \sum_{i=1}^M N_i (FP_i^* - ORF_i^*)^2}_{\text{(reliability)}} - \underbrace{\frac{1}{n} \sum_{i=1}^M N_i (ORF_i^* - SC)^2}_{\text{(resolution)}} + \underbrace{SC(1 - SC)}_{\text{(uncertainty)}}$$

$M$ : number of probability bins (normally 11)

$N$ : number of data pairs in the bin

$FP_i^*$ : binned forecast probability {0.0, 0.1,...1.0}

$ORF_i^*$ : observation for the bin {0.0=yes, 1.0=no}

$SC$ : sample climatology (total occurrences / total forecasts)

### Skill Score

$$SS = 1 - \frac{\text{reliability} - \text{resolution} + \text{uncertainty}}{0 - 0 + \text{uncertainty}} = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}$$

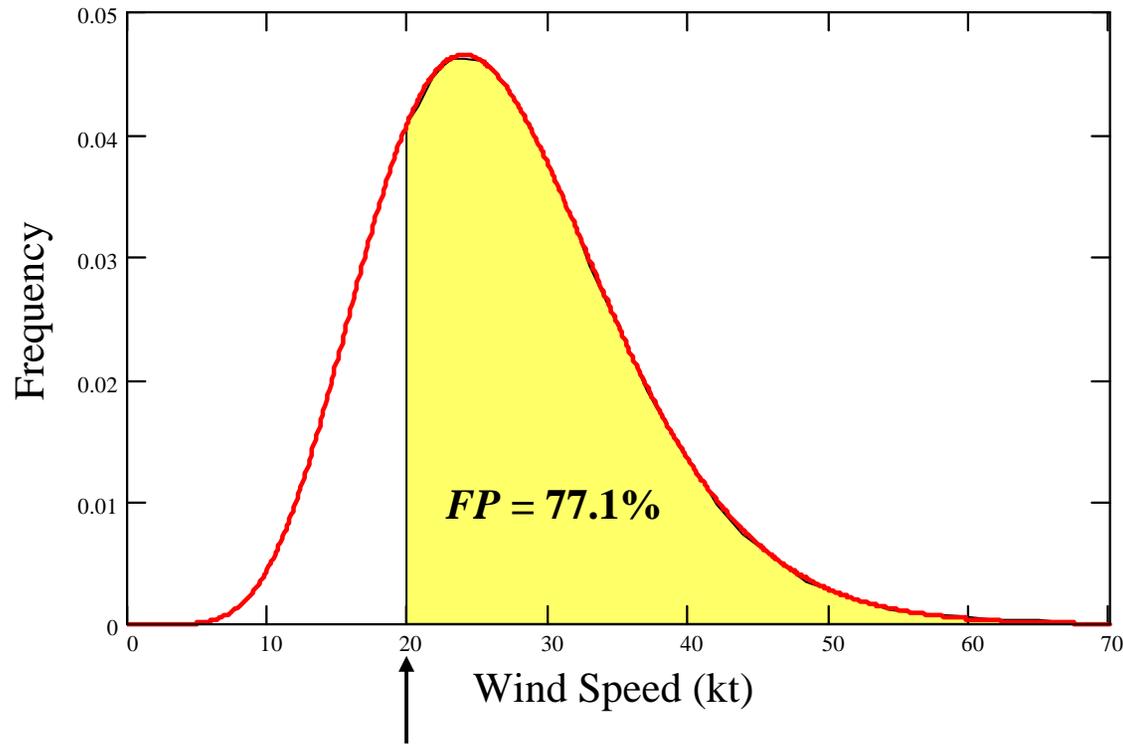
### ADVANTAGES:

- 1) No need to know long-term climatology of the parameter
- 2) Can compute  $SS$  and visualize  $BS$  in a reliability diagram

## Ideal Calculation of Forecast Probability (*FP*)

Given a very large ensemble, a PDF could be found a grid point for any parameter (e.g., wind speed,  $W_s$ ).

For a certain threshold, say  $W_s \geq 20\text{kt}$ , the *FP* is then simply the area under the PDF to the right ( $1-p$  value)



Unfortunately, we work with very small ensembles so we can't make a good estimate of the PDF. Plus, we often do not even know what PDF shape to fit.

So we are forced to estimate *FP* by other means, for a set of  $W_s$  forecasts at a point such as:

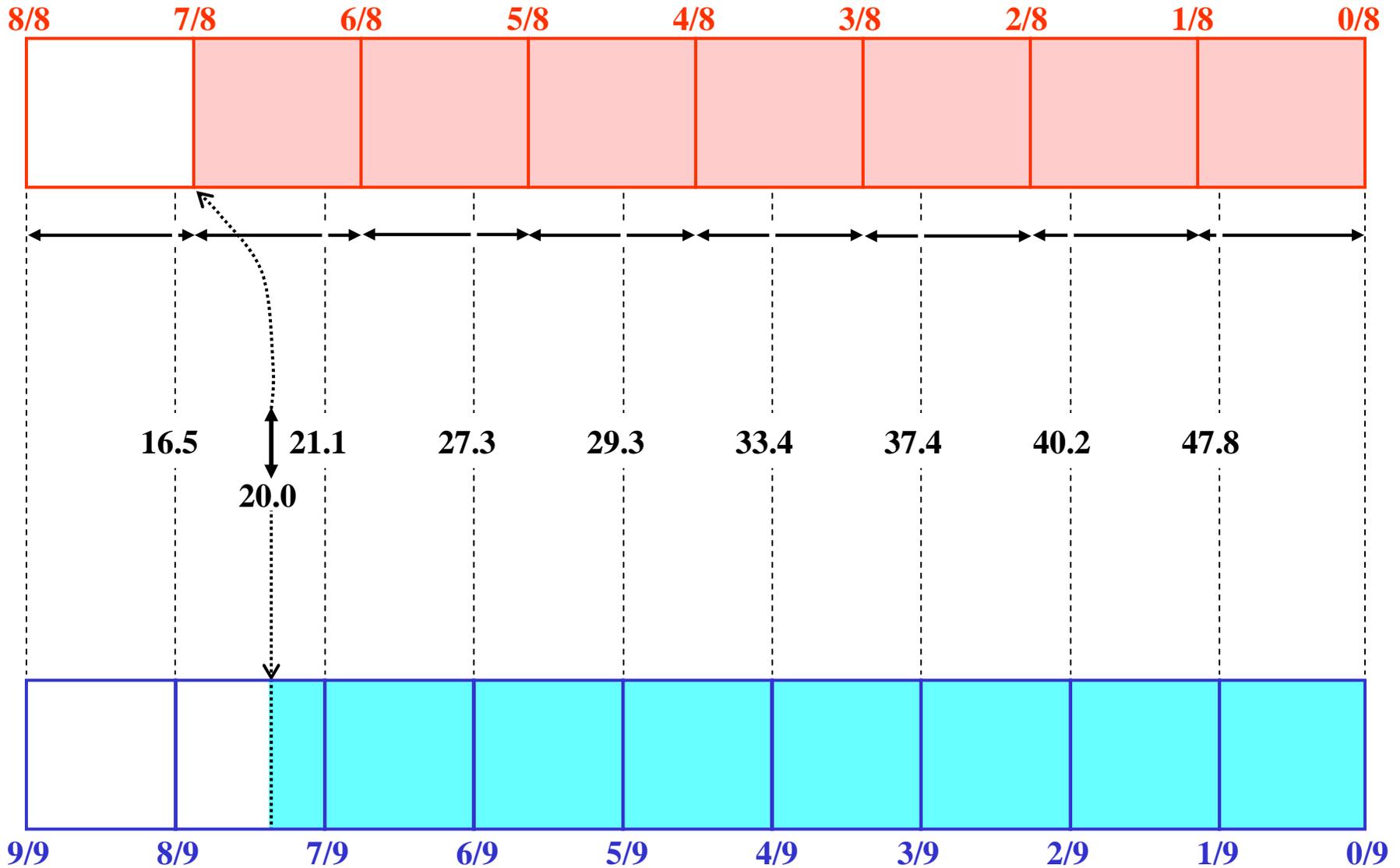
$W_s = \{16.5 \ 21.1 \ 27.3 \ 29.3 \ 33.4 \ 37.4 \ 40.2 \ 47.8\}$

Note: These are random draws from the PDF above

## Democratic Voting *FP*

“pushes” *FP* towards the extreme values, so high *FP* is normally over-forecast and low *FP* is normally under-forecast.

$$FP = 7/8 = 87.5\%$$



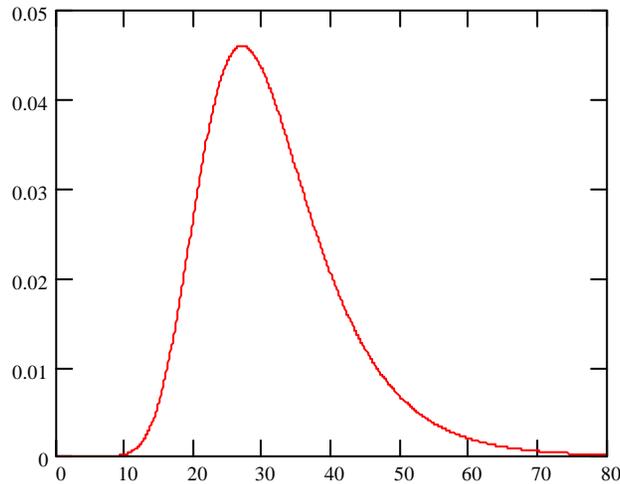
## Uniform Ranks *FP*

a continuous, more appropriate approximation.

$$FP = 7/9 + [ (21.1 - 20.0) / (21.1 - 16.5) ] * 1/9 = 80.4\%$$

# FP When Threshold Falls in an Extreme Rank

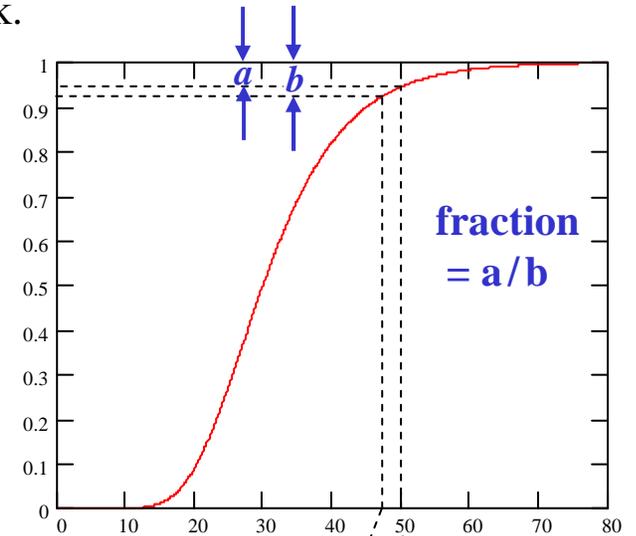
Use the tail of a Gumbel PDF to approximate the fraction for the last rank.



$$G_{CDF}(x) = \exp\left[-\exp\left(\frac{\xi - x}{\beta}\right)\right]$$

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi}$$

$$\hat{\xi} = \bar{x} - \gamma\hat{\beta}$$



0.0,  
-∞

16.5

21.1

27.3

29.3

33.4

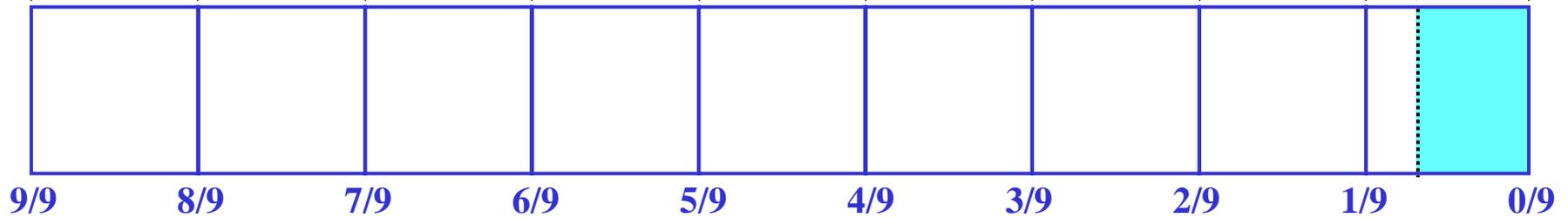
37.4

40.2

47.8

50.0

∞



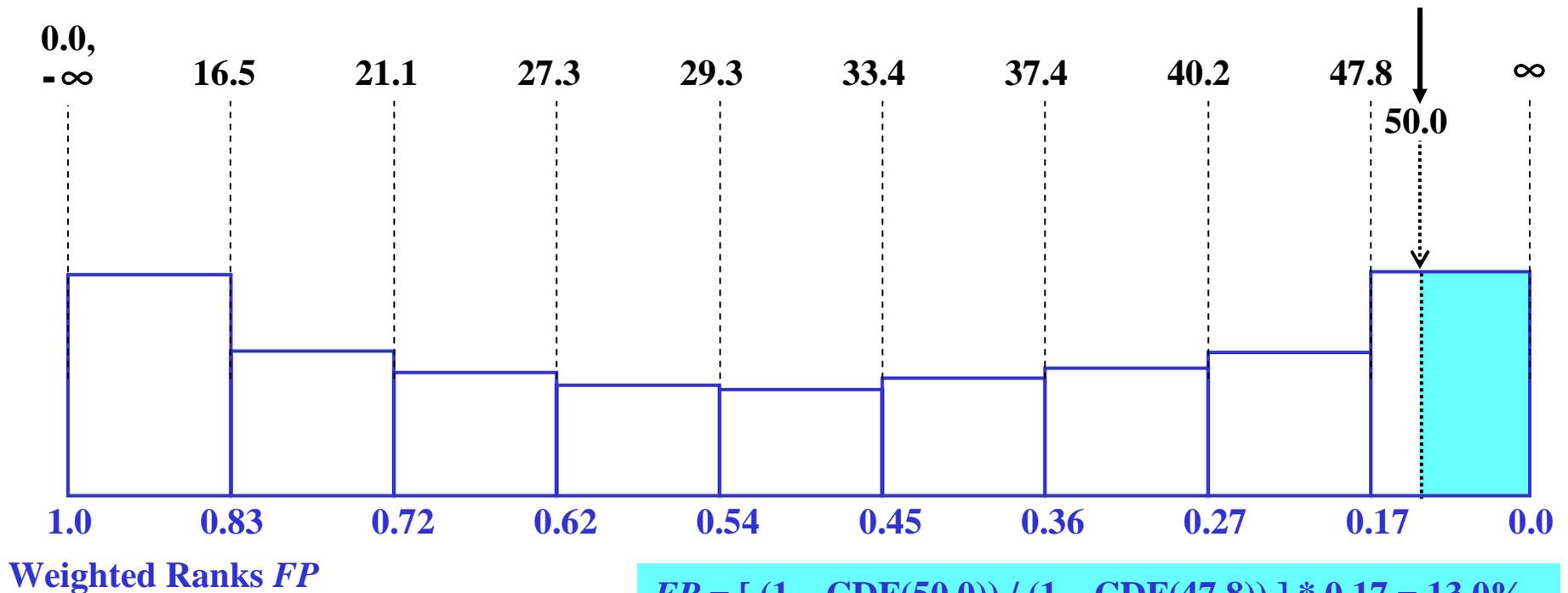
Uniform Ranks FP

$$FP = [ (1 - G_{CDF}(50.0)) / (1 - G_{CDF}(47.8)) ] * 1/9 = 8.5\%$$

# Calibration by Weighted Ranks

Use the verification rank histogram from past cases to define non-uniform, “weighted ranks”.

The ranks to sum up and fraction of the rank where the threshold falls are found the same way as with uniform ranks, but now the probability within each rank is the chance that truth will occur there.



# Uniform Ranks vs. Democratic Voting

## Data Info

- $P(\text{MSLP} < 1002\text{mb})$
- Verification: centroid analysis
- 70 forecasts (Nov 25, 2002 – Feb 7, 2003)
- Applied 2-week, running bias correction
- 36km, Outer Domain
- Lead time = 48h

